A Theory of Geographic Variations in Medical Care*

Pierre Thomas Léger † Robert J. Town ‡ Jiashan Wu §

October 11, 2019

Abstract

In this paper we provide a testable theory of geographic variations in health care expenditure and utilization that reconciles the stylized facts on geographic variations. Our model is rather straightforward yet provides insight into the underlying phenomena of variations, its potential causes and the welfare and consequences of different policy initiatives. We model imperfectly competitive, capacity constrained and perfectly altruistic providers as facing two different patient populations: privately insured and Medicare. Providers face fixed prices for treating Medicare population while they negotiate reimbursement rates for their privately insured patients. In our model, payers have different technologies for monitoring provider behavior. Unlike the widely held hypothesis that geographic differences are driven by differences in provider culture, our model focuses on differences in provider incentives that lead to differences in the care that is delivered. Specifically, in our framework, variation in health care utilization and expenditures is generated by underlying geographic variation in the model's primitives of provider market structure and productivity. These differences, in turn, lead to different incentives for physicians to treat based on the type of insurance of the patient. We then calibrate the model and run a series of counterfactual policy experiments.

Keywords: Geographic Variations, Oligopoly, Health Care.

JEL Classification:

*Preliminary and incomplete. Do not quote without authors' permission. We thank conference participants at the University of Chicago (IO Fest) and Jonathan Skinner for valuable comments.

†University of Illinois at Chicago, ptleger@uic.edu

‡University of Texas at Austin and NBER, robert.town@austin.utexas.edu

§University of Illinois at Chicago, jwu205@uic.edu

1 Introduction

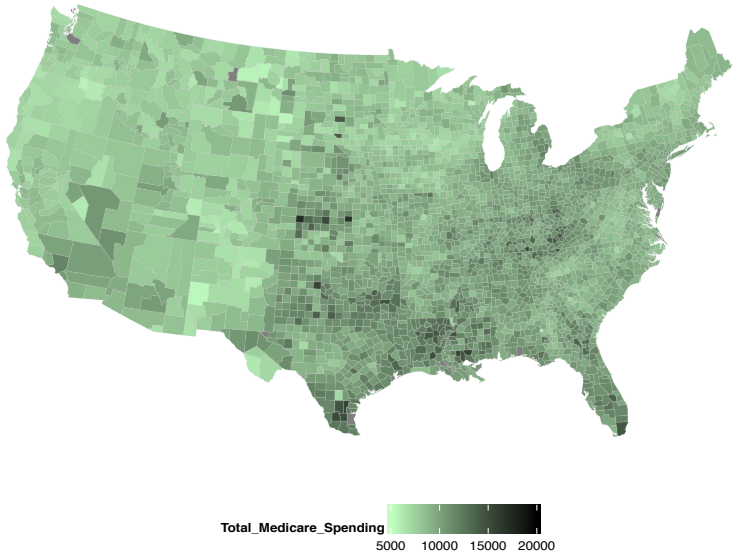
It is widely held that the US health care system is inordinately inefficient. One central piece of evidence of this inefficiency is the wide geographic variation in medical care utilization and expenditures that is not well explained by variation in underlying population health. Modern analysis of geographic variations in health care dates to Wennberg and Gittelsohn (1973) and the Dartmouth Atlas of Health Care project has spent much of the last two decades analyzing Medicare claims data and documenting this variation and its correlates.[1] Figure 1 maps the Dartmouth Atlas data on the distribution of per-capita Medicare expenditures by county for 2012. The highest expenditure counties spend well over twice the amount of the lowest spending counties. In 2013, the Institutes of Medicine (IOM) (Newhouse et al., 2013) issued an influential report documenting significant geographic variation in both the Medicare and commercially insured populations. The IOM report reiterates that much of this variation cannot be explained by demographics, observable health status and health outcomes. Furthermore, as both the IOM and the Dartmouth Atlas document, there is little correlation between health care spending and health care outcomes.[2] Clearly, given the size of the US health care market (approximately 17% of GDP), geographic variations of this magnitude likely have extremely large welfare implications. The nature of these welfare effects critically depends on the underlying mechanisms that drive this variation.

Atul Gawande's 2009 *New Yorker* article famously made geographic variation in health care cocktail party worthy discussion. He took the Dartmouth Atlas findings and focused on two neighboring cities with dramatically different average, per-capita Medicare expenditures. McAllen, TX cost Medicare nearly twice as much per beneficiary as neighboring El Paso, TX without any measurable differences in the quality of care. Building off of Dartmouth Atlas themes, Gawande points to differences in physician cultures as driving the differences

[1] Dartmouth Atlas of Health Care can be found at http://www.dartmouthatlas.org.

[2] Using a different source of identifying variation, Doyle et al. (2012) find a positive link between hospital level expenditures and patient outcomes. Our framework is flexible enough to account for either pattern through the shape of the production function.

Figure 1: Age, Sex, Price Adjusted Per-Capita Medicare Spending by County



cost in which physicians in McAllen being more profit driven than those in El Paso. The policy impact of Gawande's work is notable and played a role in inclusion of payment reform initiatives in the Accountable Care Act.[3] In this paper, we offer an alternative explanation to Gawande's.

The Dartmouth Atlas work has inspired a large, complementary body of work examining different dimensions of the geographic variations in health care.[4] One notable outcome from this body of work, in combination with the Dartmouth Atlas efforts and the IOM study, is the establishment of many stylized "facts" about the nature and pattern of geographic health care variation.[5] These facts provide important clues into the underlying mechanisms and a

[3]In 2010, Gawande was awarded AcademyHealth's Health Services Research Impact award. It is reported that Gawande's articles was required reading for White House staffers in the early years of the Obama administration.

[4]The IOM report and Chandra and Skinner (2012) provide excellent reviews.

[5]Not all studies agree on these "facts." Where there is disagreement, we use a "preponderance of the

set of predictions that a theory of geographic variations should match. In our view, the most important and interesting stylized facts are:

1. Significant variation in Medicare and privately insured spending
2. Variation in Medicare spending is driven by variation in utilization while privately insured spending variation is driven by variation in the price of care
3. Private provider prices are positively correlated with private spending and negatively correlated with Medicare spending
4. Medicare and privately insured utilization is positively *correlated* across geographies
5. Medicare and privately insured spending is positively but weakly *correlated* across geographies
6. Increases in Medicare payments *cause* privately insured reimbursements to rise.
7. Variations in Medicare and private utilization (and spending) do not translate into important corresponding variations in outcomes.

These "facts" have inspired many policy makers and commentators to make broad welfare pronouncements and policy recommendations generally focusing on the first stylized fact. For example, Wennberg et al. (2002) "propose a new approach to Medicare reform", and Fisher et al. (2009) state that the Dartmouth findings imply that policy should be "fostering the growth of more organized systems of care and implementing fundamental payment reform" (p. 852). However, both the empirical and policy sides of this literature have embarked, with some notable exceptions, without much rigorous, theoretical guidance as to the underlying causes and implications of this variation. As a consequence, the welfare and policy implications, while often extolled, are, in our view, generally not well grounded.

In this paper we provide a testable theory of geographic variations in health care expenditure and utilization that reconciles the stylized facts on geographic variations listed above. Our model is rather straightforward yet provides insight into the underlying phenomena of variations, its potential causes and the welfare consequences of different policy initiatives. Building on McGuire and Pauly (1991), we model imperfectly competitive, capacity constrained and altruistic providers as facing two different patient populations: privately

evidence" standard which combines the number of papers and the quality of the analysis/data.

insured and Medicare. Providers face fixed prices for treating Medicare population while they negotiate reimbursement rates for their privately insured patients. In our model, payers have different technologies for monitoring provider behavior.

Unlike Gawande (2009)'s hypothesis that geographic differences are driven by differences in provider culture, our model focuses on differences in provider incentives that lead to differences in the care that is delivered. Specifically, in our framework, variation in health care utilization and expenditures is generated by underlying geographic variation in the model's primitives of provider market structure and productivity. These differences, in turn, lead to different incentives for physicians to treat based on the type of insurance of the patient. Recent work has shown that geographic variation in provider productivity and market structure are empirical realities (Chandra et al. (2013), Skinner and Staiger (2007), Chandra et al. (2013), Gaynor et al. (2014)). While our theory is not complicated, the underlying variation in productivity and provider market structure interact in complex ways in the model allowing it to match the patterns of variation we observe. That is, without the model, the underlying variation in productivity or market structure cannot alone account for the observed variations behavior. Our model emphasizes that understanding geographic variation requires accounting for provider incentives that lead to linkages and spillovers between the administer price environment of Medicare and privately insured price sectors.[6,7]

The intuition behind our results is rather straightforward. Given fixed capacity, physi-

[6]Three notable exceptions to the observation that much of the study of health care variations is atheoretical are Chandra and Staiger (2007), Chandra and Skinner (2012) and Clemens and Gottlieb (2017). Chandra and Staiger (2007) construct a Roy model of productivity spillovers and show that in equilibrium this model will generate variations in treatment intensity by affecting the patient severity cutoff rule for determining the intensity of treatment. Their model can explain several patterns in the treatment of acute myocardial infarction patients in the early 1990s. However, their model cannot readily account for the weak correlation between Medicare and privately insured populations health care expenditures or the observation that price variation drives much of the variation in private health care expenditures.Chandra and Skinner (2012) show that geographic variation in utilization can be driven by exogenous variation in productivity, reimbursements, patient out-of-pocket expenditures or malpractice risk. However, they do not attempt to reconcile the patterns of variation with their theory. The work that is most similar to ours is by Clemens and Gottlieb (2017). Like us, they model linkages between the Medicare and commercially insured population. Their primary focus is on the pricing links between the two sectors while we are attempting to explain a broader variety of facts surrounding geographic variation.

[7]We provide a more comprehensive review of both the theoretical and empirical literature on geographic variations below.

cians must decide how to allocate their efforts between Medicare and private pay patients. Physicians allocate that effort based on the returns they receive which, in turn, depend upon the relative reimbursement rates (which is a function of market power for privately insured patients), their productivity and the value they place on patient welfare. We model physician-insurer negotiation explicitly to characterize the role of market power on physician-insurer-region specific reimbursement rates.

Our model has sharp implications...

Our model is then calibrated to fit data from both Medicare and commercial markets. More specifically, we specify all the functional forms and propose several appropriate distributions based on the theoretical model. We then derive the closed forms for all the moments of interests. By matching the derived moments with empirical data, we manage to solve for proper values of all of the model's parameters. The solved parameters are then applied to our stimulation exercise.

[Insert Here: Findings from the calibration exercise]

The rest of the paper is organized as follows. Stylized facts and a review of the theoretical literature are presented in Section 2. Section 3 presents the theoretical model. Empirical and policy implications of the model are presented in Section 4. In Section 5, we calibrate our model to match a series of moments and run counterfactual policy experiments. Conclusions are drawn in Section 6.

2 Stylized Facts and Theoretical Literature

2.1 Stylized Facts of Geographic Variations

The observation that health care providers do not treat like patients similarly dates to the 1938 study of tonsillectomies in England (Glover, 1938). The modern analysis of variations starts with Wennberg and Gittelsohn (1973) who shows that there is variation in medical care among providers in a small geographic area. Over the last 30 years, a large literature has

documented and examined the variation in medical care expenditures.[8] Much of this work has been done by researchers under the umbrella of the Dartmouth Atlas of Health Care project. The first volume was published in 1996 (Wennberg and Cooper, 1996). The impact of the Dartmouth work had on the health service research and health policy community is monumental inspiring many other researchers to jump in to provide more color to the analysis. The policy implications of variations are significant leading to reports by CMS, MedPAC and and most recently the detailed IOM report.

One of the important aspects of this literature is the focus on analyzing detailed medical claims data for millions of patients. The quantity and quality of the data allow for a detailed examination of geographic variations. Below we reproduce the major, important aggregate patterns that have been documented in this literature. To do this, we use publicly available data from the IOM project. These data provide aggregated measures of spending, utilization and quality of care by insurance type for 306 hospital referral regions (HRR) using millions of Traditional Medicare and privately insured patient claims. HRRs, which were created under the Dartmouth Atlas project, are agglomerations of zip codes that approximate regional hospital markets. These agglomerations are based on hospital admission travel patterns.[9]

The IOM presents several versions of the expenditure data and we use the version that has been adjusted for age, sex, race, input price and health status. These adjustments likely understate the true variations as the health status adjustments are based on claims data and likely capture some supply side coding behavior. Also, the IOM used two distinct source of commercial claims data with two different teams performing the analysis. We combine those two data sources into a weighted (by sample size) average of expenditure and utilization for each HRR.

[8]More complete reviews of this literature can be found in Skinner (2011), Chandra and Skinner (2012) and (Newhouse et al., 2013).

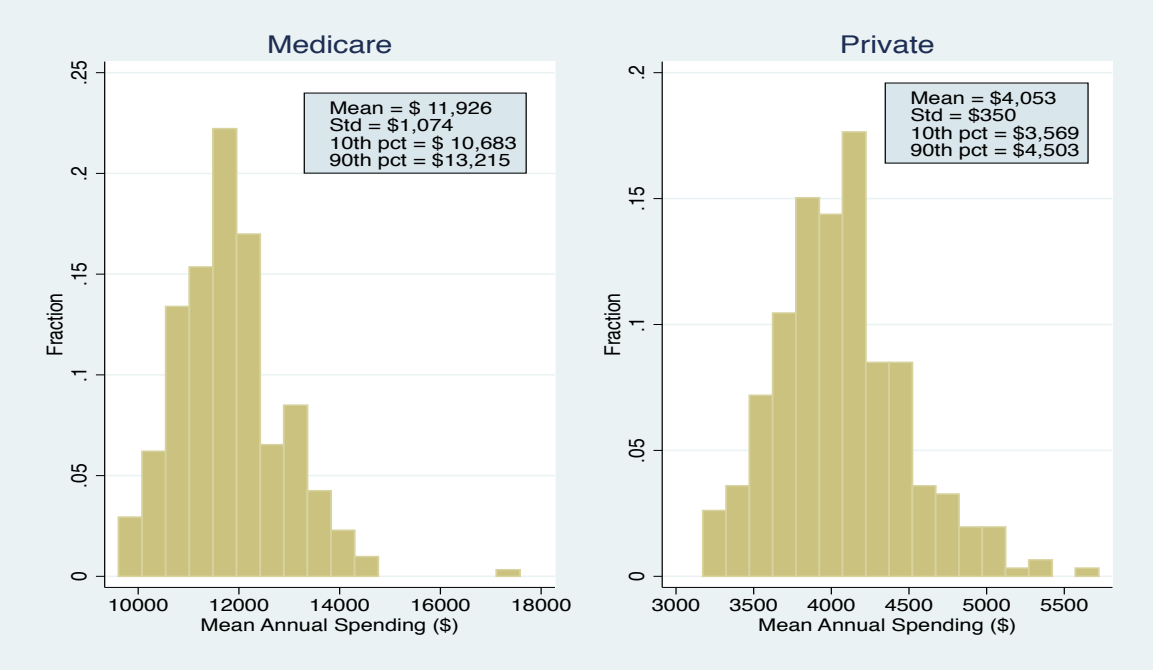
[9]More detail on HRR construction can be found here: http://www.dartmouthatlas.org/data/region.

Stylized Fact 1: There is significant geographic variation in Medicare and privately insured spending that is not explained by variation in health status and outcomes

The Dartmouth Atlas popularized the notion that geographic variations in health care spending is an important, policy relevant phenomenon. Figure 2 displays histograms of the Medicare and commercial spending by HRR for 2009. The variation for both populations is meaningful. The ratio of the 90th to 10th percentile spending is 1.24 and 1.26 for the Medicare and commercially insured populations, respectively. The degree of variation in these IOM data are notably less that the variation in the Dartmouth Atlas data presented in the Introduction which is likely a result of the health status and other demographic adjustments. The finding of variation of this magnitude in both the Medicare and commercial populations is extraordinarily robust across numerous studies spanning time, data sources and methodological approaches. Fisher et al. (2003a), Fisher et al. (2003b), MedPAC (2011), Chernew et al. (2010), Philipson et al. (2010), McKellar et al. (2014) and Romley et al. (2014) all document significant geographic variation in health care expenditures and utilization. This variation is also significant from a policy perspective. Moving mean expenditures in HRRs above the median to the median would reduce total Medicare fee-for-service expenditures by 4% or $60 billion annually.

However, this finding of significant geographic variation is not without its critics. The primary issue is whether this variation is driven by demand side factors (e.g. health status, income) which is less interesting for economics and policy or whether this variation is driven by supply side behavior of hospitals and physicians. For example, Sheiner (2014) argues that socioeconomic factors that affect the need for medical care, as well as interactions between the Medicare system and other parts of the health system, accounts for most of the variation in Medicare health spending. This study, in turn, has its critics and the broad base of evidence aligning with the Dartmouth Atlas view. Using data on Medicare enrollee migration, Finkelstein et al. (2016) report that 50 to 60% of the variation in Medicare expenditures are

Figure 2: Variation in Medicare and Commercial Spending



attributable to supply-side factors. Examining physician responses to vignettes, Cutler et al. (2013) find little evidence that the geographic variation is driven by demand-side factors.

The finding of variation holds even when drilling down more fine characterizations of the care. Variation has been found for end-of-life care, back surgery, hip replacement and knee surgery in the Medicare population. The IOM study found variation for inpatient, outpatient, emergency department visits, pharmaceutical utilization and use of imaging services.

Stylized Fact 2: Geographic variation in Medicare spending is driven principally by variation in utilization while privately insured spending variation is primarily driven by variation in the price of care

An obvious question raised by the findings of Stylized Fact 1 is whether the variation in expenditures is driven by variation in prices, quantities or both. Medicare prices are administratively set suggesting that all Medicare variation must be driven by variation in quan-

tities. However, the formulas determining those rates can vary by provider and geography. Furthermore, through differences in coding practices providers are able to manipulate effective Medicare reimbursement rates (McClellan (1997), Dafny (2005), Silverman and Skinner (2004)). Thus, it is an open question whether Medicare spending variation is driven by prices or quantities.

The IOM report decomposed the variation in both Medicare and privately insured patients. They concluded that variation in Medicare spending is driven primarily by variation in services delivered. For the privately insured, variation in spending was principally driven by variation in prices implying that utilization across geographies. Specifically, variation in the price of care accounted for 70% of the variation in the privately insured spending. The IOM's work builds upon the analysis in Gottlieb et al. (2010) who also found that Medicare spending variation was driven by utilization and not prices.

Stylized Fact 3: Private provider prices are positively correlated with private spending and negatively correlated with Medicare spending

Medicare and privately insured sectors likely do not operate in isolation from one another. In fact, one of the central features of our theory is that as long as providers face capacity constraints, these sectors are linked. Empirical evidence of this linkage is reported in Romley et al. (2014). There they examine the relationship between the mean HRR price for medical care provided to privately insured patients and private and Medicare spending. These correlations are striking. Increases in private prices are associated with increases in private spending and a decreases in Medicare spending.

Stylized Fact 4: Medicare and privately insured utilization is positively *correlated* across geographies

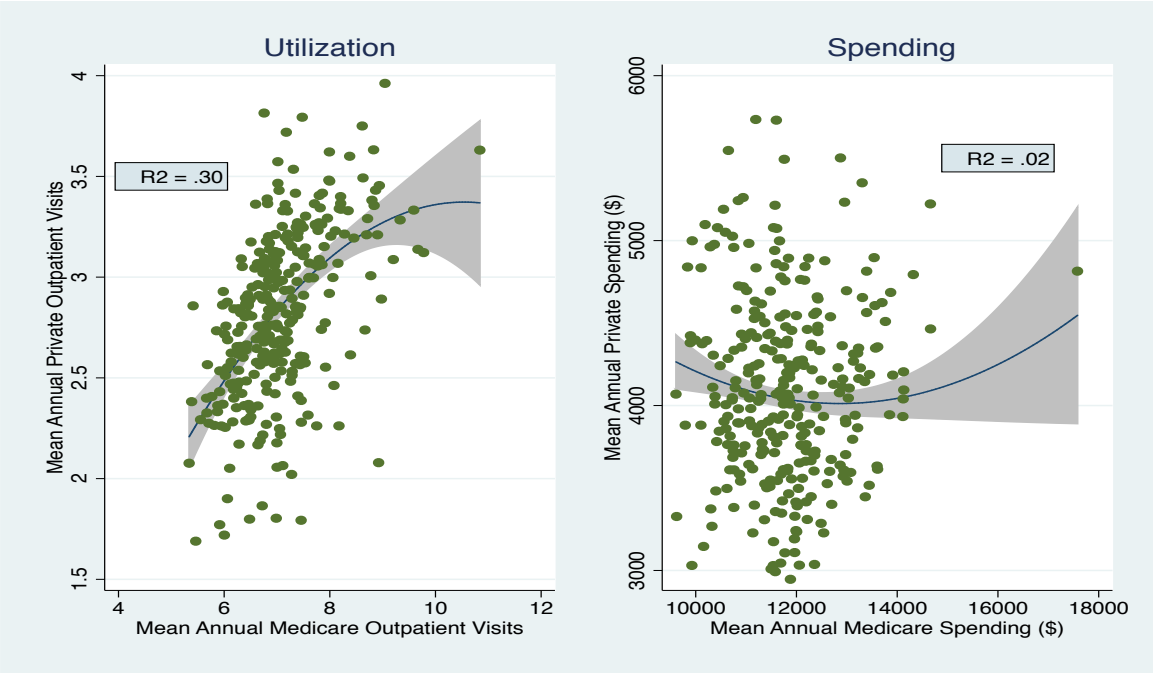
Stylized Fact 5: Medicare and privately insured spending is *uncorrelated* to weakly *correlated* across geographies

[***Modify to reflect new graphs with same unit of observation***]Correlations between utilization and spending between Medicare and the privately insured provide insights into the nature of the spillovers between the sectors. In Figure 2.1 we present these correlations. In the first panel we present the utilization (based on number of outpatient visits) scatterplot and the quadratic fit regression line which visually present the findings in Stylized Fact 4. The correlation in utilization between Medicare and the private sectors is positive and strong. The R^2 on the regression is .30. In the second panel we present the same analysis for spending. Here the story is very different. There is very little correlation between per-capita spending in the Medicare and commercial population. This correlation pattern is similar to the one reported in Chernew et al. (2010).

Stylized Fact 6: Increases in Medicare payments *cause* privately insured reimbursements to rise

There is a large literature examining the relationship between public payer reimbursement rates and the rates private payers negotiate with providers. The common view is that 'cost-shifting' occurs. In this review of the literature, Frakt (2011) finds little credible evidence for 'cost-shfiting,' the idea that providers offset decreases in public reimbursement rates by increasing private rates. More recently, relying on well identified variation in Medicare payments, Clemens and Gottlieb (2017) find that increases in Medicare reimbursements lead to an increase in private reimbursement rates. They find that the correlation is strongest when insurers are concentrated, small physician groups, and in more competitive physician markets.

Figure 3: Correlations in Medicare and Commercial Spending and Utilization



Stylized Fact 7: Variations in utilization (and spending) are not consistently correlated with health outcomes

Health outcomes also vary by geography (e.g. Finkelstein et al. (2018) and literature cited therein). However, the variation in health outcomes appears to be unrelated to variation in healthcare spending and utilization (e.g. Hussey et al. (2013), Skinner (2011), Fisher et al. (2003b), Fisher et al. (2003a)). For the private market, it is not surprising that spending variation does not explain health outcomes given Stylized Fact 2 which describes that most of the variation in private spending is a consequence of variations in prices (not utilization).

The observation that health care outcomes are not positively correlated with health care spending has lead to two types of explanations. The first and most common explanation is "flat of the curve" medicine (e.g. Fuchs (2004)) whereby additional inputs in a standard concave production function have decreasing marginal product. The second explanation is that, for a number of possible reasons, the productivity of care may differ across geographies and

this productivity may be correlated in complex ways with input use (e.g Chandra and Staiger (2007)). Our model will allow for both types of explanations for this lack of correlation.

2.2 Theoretical Analysis of Geographic Variations

No single model considers or explains all, or even most, of the stylized facts we list above. Nonetheless, several models of geographic variations have been developed and we discuss them below.

Skinner (2011) builds a model that seeks to explain the geographic variation in spending and utilization for *Medicare* patients. To do so, he provides a two-period model where individuals seek medical care from partially altruistic physicians and follow their recommendations. In his model, individuals value medical consumption uniquely as a means to augment their second-period utility. He shows that variations in Medicare spending across different geographical areas can be generated by both demand and supply side factors.[10] Because Skinner treats the Medicare market in autarky (i.e., does not consider interdependencies across markets), his model does not speak to most of the stylized facts spelled out above.

Philipson et al. (2010) provide a more macro-level model that assumes (rather than generates) within-market variation across physicians in their provision of care (i.e., a market specific distribution of care y across physicians is given by $F(y)$). Furthermore, their model assumes (again, rather than generates) that the distribution of care $F(y)$ varies from one market to the next. That is, they take as given across-market variation in utilization (and spending under a common price). Finally, the authors assume that physicians treating privately insured patients are subject to an exogenously determined "hard" utilization review which is also assumed to be common across markets. More specifically, the authors assume a hard constraint on y such that $F(y)$ distributions (along a common support) are

[10] Among demand-side factors are: (i) differences in health status and need, (ii) differences in income, (iii) differences in prices paid (or insurance generosity), (iv) differences in patient preferences, and (v) differences in access (or other constraints). Among supply-side factors are: (i) differences in practice norms (education, training,...), (ii) differences in ability/productivity, (iii) differences in altruism/preferences (i.e., how physicians value patient health relative to financial gains), and (iv) differences in financial incentives and competition.

right-truncated at a common y^{ur}. Given these assumptions, two basic results (i.e., testable hypotheses) follow. First, the within-market variance in utilization amongst private payers is smaller than the within-market variance in utilization amongst Medicare beneficiaries. This result mechanically falls out by simply truncating the private care distribution. Second, the difference in within-market variance in utilization across private and Medicare patients is likely to increase as the average Medicare utilization increases. This also again mechanically falls out by the single right-sided censoring at y^{ur}. They bring these two theoretical results to the data and find evidence to confirm them. Philipson et al. (2010) say little about why different regions face different distribution of care patterns nor why the private sector restrictions would be both "hard" in nature and common to all geographic regions. There model cannot, however, speak to endogenous private-sector price setting nor the many interdependencies across markets in utilization and spending.

3 Theoretical Model of Service Provision

In this section, we build a model where the relationship that exists between the private and Medicare markets is endogenous to the market structure.

3.1 A simple model of service provision

Consider a partially altruistic physician j, in region s, who faces two types of patients: those covered by Medicare (M) denoted by i and those covered by private insurance (P) denoted by k. Further assume that the physician receives a fee F^M and $F^P_{j,k,s}$ per quantity of care provided to Medicare and private-payers, respectively. The Medicare fee (F^M) is assumed common across physicians and regions to reflect administrative nature of the fees, while the private fee ($F^P_{j,k,s}$) is assumed to be physician j, patient k (i.e., private insurance) and region

s specific to reflect the negotiated nature of such fees.[11,12]

Let physician j's utility of providing q^M units of care to a Medicare patient i who suffers from illness severity θ_i^M in region s be given by:[13]

$$U_{j,i,s}^M = h(\theta_i^M, q_i^M) + F^M q_i^M - c(q_i^M), \tag{1}$$

where $h(\theta_i^M, q_i^M)$ represents patients' health production function (common across patients) with typical assumptions of $h'_{\theta^M} < 0$, $h'_{q^M} > 0$, $h''_{q^M} < 0$ and where $c'(.) > 0$ and $c''(.) > 0$.

Similarly, let physician j's utility of providing q^P units to a private patient k who suffers from illness severity θ_k^P in region s be given by:

$$U_{j,k,s}^P = h(\theta_k^P, q_k^P) + F_{j,k,s}^P q_k^P - c(q_k^P). \tag{2}$$

Now, in order to reflect some additional features of the market, we augment the above setup in several ways. First, we assume that physicians face a physician-region specific capacity constraint in the total quantity of care units ($\overline{Q}$) which is increasing in the physician's productivity: α_j - where α_j is drawn from a region specific distribution which we characterize below. More specifically, where $\sum_{k=1}^K q_k^P + \sum_{i=1}^I q_i^M \leq \overline{Q}(\alpha_j)$ and where $\overline{Q'}(\alpha_j) > 0$.

Furthermore, recent work has found that measures on physicians bargaining leverage are correlated with the fees they receive from private insurers (McKellar et al. (2014), Dunn and Shapiro (2014), Kleiner et al. (2015) and Gaynor et al. (2014)). We incorporate this feature into our model by allowing physicians to vary in their bargaining position which affects the fees they receive from the private insurer. We denote the physician bargaining leverage as μ_j and the reduced-form relationship between the private fee and bargaining

[11] We omit the s subscript on the Medicare fee (F^M) as are administratively set and where: (i) all providers within a geography s are subject to the same fee and (ii) differences in Medicare fees across geographies are do to variations in cost of living.

[12] In the model, we assume that each patient k is associated with a particular insurance provider, and thus, k denotes both the private patient and their insurance provider.

[13] Unlike Skinner (2011), we assume a paternalistic form of altruism - i.e., the physician cares about the patient's health not the patient's demand for care. For insured patients, these are essentially equivalent.

leverage as $F^P_{j,k,s}(\mu_j)$, where F is weakly increasing in μ_j (i.e., $F'^P_{j,k,s}(\mu_j) \geq 0$). We endogenize the relationship between the physician's bargaining power μ_j and the physician's private fee $F^P_{j,k,s}$ explicitly in Section 3.2 below.[14]

As alluded to above, an important feature of private insurance relative to Medicare is that they have the incentive to more closely monitor physician behavior to reduce physician agency. In some important sense, this is one of the principle functions of managed care. We incorporate this into the model by assuming that private insurance providers engage in cost controls in the form of financial penalties for 'excessive' costs $f(F^P_{j,k,s}(\mu_j)q^P_k)$ where $f'_{F^P} > 0$ and $f'_{q^P} > 0$.[15]

For simplicity and without loss of generality, we assume that each physician has two patients: one Medicare and one privately insured. We additionally assume that the care provided to private and Medicare patients is homogenous in terms of its costs. That is, the marginal cost of providing a unit of care is invariant to the type of patient served.

Assuming that: (i) the capacity constraint is binding, (ii) physicians maximize utility over their entire (i.e., two patient) population (i.e., follow a utilitarian rule), and (iii) physicians value their Medicare and private patients' health equally, yields the following physician-j's, in region s, maximization problem: [16]

$$\begin{aligned} max\mathcal{L}_{q^M,q^P,\lambda} = h(\theta^M_i, q^M_i) + F^M q^M_i + h(\theta^P_k, q^P_k) + F^P_{j,k,s}(\mu_j)q^P_k \\ -c(q^M_i + q^P_k) - f(F^P_{j,k,s}(\mu_j)q^P_k) + \lambda(\overline{Q}(\alpha_j) - q^M_i - q^P_k). \end{aligned} \tag{3}$$

The FOC conditions with respect to q^M, q^P and λ are respectively given by:

$$h'_2(\theta^M_i, q^M_k) + F^M - c'(q^M_i + q^P_k) - \lambda = 0, \tag{4}$$

[14]Bargaining leverage, bargaining power and market power are used interchangeably throughout.

[15]Town et al. (2011) provide a theoretical model the relationship between physician market structure and the private payer contract structure. Consistent with our result, they find that physician with higher bargaining leverage have "lower power" contracts.

[16]If the capacity constraint were not binding, then how much the physician values her patient's health relative to her net income would come into play.

$$h_2'(\theta_k^P, q_k^P) + F_{j,k,s}^P(\mu_j) - c'(q_i^M + q_k^P) - f_{q^P}'(F_{j,k,s}^P(\mu_j)q_k^P) - \lambda = 0, \tag{5}$$

$$\overline{Q}(\alpha_j) - q_i^M - q_k^P = 0. \tag{6}$$

Together, the first-two FOCs yield:

$$h_2'(\theta_i^M, q_i^M) + F^M = h_2'(\theta_k^P, q_k^P) + F_{j,k,s}^P(\mu_j) - f_{q^P}'(F_{j,k,s}^P(\mu_j)q_k^P). \tag{7}$$

Finally, by substituting the capacity constraint (which we assume is binding) into this last equality, the optimal amount of q^P satisfies:

$$h_2'(\theta_i^M, \overline{Q}(\alpha_j) - q_k^P) + F^M = h_2'(\theta_k^P, q_k^P) + F_{j,k,s}^P(\mu_j) - f_{q^P}'(F_{j,k,s}^P(\mu_j)q_k^P). \tag{8}$$

This last condition implicitly defines physician-j's region-s specific supply curve of care to the private patient k (with corresponding private insurance) for the given pair of illness severities (θ_i^M and θ_k^P) and for a given set of parameters α_j, μ_j. Coupled with the capacity constraint, it also implicitly defines physician-j's region-s specific supply curve for care to the Medicare patient for the same pair of illness severities and parameters. Analogously, one can derive physician-$j's$ region-s specific supply curves of care to the private and Medicare patients (respectively) across all potential illness severity pairs and possible parameter values. We denote physician-j's region-s specific Medicare and private-payers supply functions as $q_{j,i,s}^M = \mathcal{Q}(F^M, F_{j,k,s}^P, \theta_k^P, \theta_i^M; \mu_j, \alpha_j)$ and $q_{j,k,s}^P = \mathcal{Q}(F^M, F_{j,k,s}^P, \theta_k^P, \theta_i^M; \mu_j, \alpha_j)$, respectively.

Under standard conditions, q_i^M (q_k^P) is increasing (decreasing) in θ_i^M and decreasing (increasing) in θ_k^P. Thus, the underlying need for care in one patient population can affect the provision of care to the other (see Appendix 1 for proof and conditions). Furthermore, holding all other elements constant, q_i^M and q_k^P are increasing in the physician's productivity, or equivalently, capacity constraint but where the effect on Medicare is greater than private

utilization under reasonable conditions ($\frac{dq^M}{d\overline{Q}(\alpha)} > \frac{1}{2}$ and $0 < \frac{dq^P}{d\overline{Q}(\alpha)} < \frac{1}{2}$)(see Appendix 2 for proof and conditions).

We denote the optimal level of care supplied by physician j in region s to her Medicare and private patients i and k who suffer from θ_i^M and θ_k^P (with given prices (F^M, $F_{j,k,s}^P$) and parameters (α_j, μ_j)) as $q_{j,i,s}^{*M}(\theta_i^M, \theta_k^P)$ and $q_{j,k,s}^{*P}(\theta_i^M, \theta_k^P)$, respectively. Finally, we denote the difference in the two optimal quantities as $\Delta_{j,k,i,s}(\theta_i^M, \theta_k^P)$=$q_{j,k,s}^{*P}(\theta_i^M, \theta_k^P)$-$q_{j,i,s}^{*M}(\theta_i^M, \theta_k^P)$. In Appendix 1, we also show that $\Delta_{j,k,i,s}(\theta_i^M, \theta_k^P)$ is increasing in θ_k^P and decreasing in θ_i^M. It is worth noting that these upward sloping supply curves become completely inelastic at the capacity constraint. For the rest of the exercise, we will assume interior solutions.

3.2 Endogenous private fees

In this section we model the physician-private insurer price setting process which was previously represented by its reduced-form relationship: $F_{j,k,s}^P(\mu_j)$ and $F_{j,k,s}^{\prime P}(\mu_j) > 0$.

The physician's problem:

Consider, as above, that the physician is endowed with both a private (served by a particular insurance provider) k and a Medicare i patient and that the private ($F_{j,k}^P$) and Medicare (F^M) reimbursement rates are given (while omitting the s subscript for compactness).

The physician's utility when treating k and i (with illnesses θ_i^M and θ_k^P) optimally is given by:

$$U_{j,k,i}^A \equiv h(\theta_i^M, q_i^{*M}) + F^M q_i^{*M} + h(\theta_k^P, q_k^{*P}) + F_{j,k}^P q_k^{*P} - c(\overline{Q}(\alpha_j)) - f(F_{j,k}^P q_k^{*P}), \tag{9}$$

where $U_{j,k,i}^A$ constitutes the physician's agreement utility at given reimbursement rates and where q_i^{*M} and q_k^{*P} denote the equilibrium quantities.

Next assume that if the physician refuses to treat her private patient k (with corresponding reimbursement rate $F_{j,k}^P$), she can replace him or her with another private patient k' or another Medicare patient i' with probabilities μ_j and $(1-\mu_j)$, respectively.[17] So μ_j, which we defined

[17]Where a new private patient k' would be associated with an expected (equilibrium) reimbursement rate

as the physician's market power, is given a specific meaning here. That is, a physician with greater market power is one who can more easily replace a current patient with a private-paying one.

Thus, if the physician refuses to treat patient k, her expected utility is given by:

$$U^D_{j,k',i'} \equiv \mu_j[h(\theta^M_i, q^{**M}_i) + F^M q^{**M}_i + h(\theta^P_{k'}, q^{**P}_{k'}) + F^P_{j,k'} q^{**P}_{k'} - f(F^P_{j,k'} q^{**P}_{k'})]$$
$$+(1-\mu_j)[h(\theta^M_i, q^{***M}_i) + F^M q^{***M}_i + h(\theta^M_{i'}, q^{***M}_{i'}) + F^M q^{***M}_{i'}] - c(\overline{Q}(\alpha_j)) \quad (10)$$

where $U^D_{j,k',i'}$ constitutes the physician's (expected) disagreement utility and where (**) denote the new equilibrium quantities when the physician draws another private patient while (***) denote the new equilibrium quantities when the physician draws another Medicare patient. We make the simplifying assumption that the private fee associated with treating patient k' as given. For simplicity, we assume that a new Medicare patient would have the same illness shock as a new private patient.

Two things are worth noting here:

1. By comparing the "agreement" and "disagreement" utilities, and assuming that $F^P_{j,k'} > F^P_{j,k}$, one can see that there exists a minimum $F^P_{j,k}$, denoted $F^P_{min,j}$, such that the physician is willing to continue to treat her patient k (i.e., the agreement utility is greater than the disagreement utility) rather than try her luck on a new patient i' or k', and [18]

2. Again, assuming that $F^P_{j,k'} > F^P_{j,k}$, the minimum fee $F^P_{min,j}$ is increasing the physician's μ_j (i.e., the more likely the physician is to gain a private patient rather than a Medicare patient if she were to drop her current k patient, the more she requires a larger private fee to continue to treat her current private patient k).

The insurance provider's problem:

Next, assume that the insurance provider's utility/profit is increasing in its patient k's health

$F^P_{j,k'}$.

[18] Assuming, for simplicity, that patients i' and k' have the same illness severities as the k patient they potentially replace.

(which is itself increasing in quantity q_k^{*P}) but decreasing in the price it pays the physician ($F_{j,k}^P$). If the insurance provider offers a private fee above the physician j's minimum fee (i.e., $F_{j,k}^P > F_{min,j}^P$), then the physician will keep her current private patient k and respond with the corresponding optimal quantity q_k^{*P}. Insurance provider k's profit is then given by:

$$V_k^A \equiv V(h(\theta^P, q_k^{*P}), F_j^P q_k^{*P}) \tag{11}$$

where $V_1' > 0$ and $V_2' < 0$. This characterizes the physician's agreement profit.

If, however, the insurance provider offers a fee below physician j's minimum fee (i.e., $F_{j,k}^P < F_{min,j}^P$), the physician will refuse to continue to see patient k and what the insurance provider will receive is normalized to zero:

$$V_k^D \equiv 0. \tag{12}$$

This characterizes the insurance provider's disagreement profit.

The physician-insurer reimbursement-setting process:

We next consider the physician insurance-provider interaction in a sequential framework. First, we assume that the insurance provider makes a take-it-or-leave-it offer of $F_{j,k}^P$ to physician j, which she can accept or refuse (under a full information framework).

The insurance provider will choose to offer a fee $F_{j,k}^{*P} > F_{min,j}^P$ to physician j that maximizes its agreement utility (V^A) if doing so yields more profits than offering: (i) exactly $F_{min,j}^P$, and (ii) offering less than $F_{min,j}^P$ and receiving the disagreement utility (V^D) which is normalized to zero. Similarly, the insurance provider will choose to offer exactly the minimum fee $F_{min,j}^P > F_{j,k}^{*P}$ (although greater than would be offered in the absence of a physician's ability to drop the patient k) if doing so yields greater profits than $V^D = 0$, and offering more than $F_{min,j}$ with corresponding physician response. As a result, the greater j's disagreement utility, which is itself increasing in μ, the greater will be the minimum fee $F_{min,j}^P$ that she must be offered for her to continue to treat her patient k. Thus, the fee offered to the physician

$(F^P_{j,k})$ in this take-it-or-leave-it setting is weakly increasing in physician j's minimum fee $F^P_{min,j}$ which is itself strictly increasing in probability μ_j. We represent this price-setting environment with the reduced form relationship between this likelihood (henceforth referred to as the physician's market power or bargaining leverage) and the physician's equilibrium reimbursement rate as $F^P_{j,k}(\mu_j)$.

4 Implications of the Model

We next reconsider our model in order to evaluate if and under what conditions it is consistent with the previously described Stylized Facts 1 through 7. Before deriving the model's predictions, which will yield testable empirical implications, we make the following simplifying assumptions. The capacity constraint (i.e., productivity) parameter α_j and the bargaining leverage parameter, μ_j, are drawn from a region s-specific distribution $\Gamma(\alpha, \mu; \omega_s)$ where ω_s completely characterizes the distribution Γ and $\Gamma(\alpha, \mu; \omega_s) = \Gamma(\alpha, \mu; \omega'_s)$ iff $\omega_s = \omega'_s$. The hyper-parameter ω_s is drawn from the non-degenerate distribution $K(\omega_s)$. Thus, the distribution of two of the primitives of the model (α and μ) vary by geographic regions s. The theoretical and calibration exercise below will allow us to identify the across-region relationship between capacity constraint and productivity parameters α and μ.

Before proceeding, we develop some additional notation. Let $l^z_s, z \in \{M, P\}$ denote the mean Medicare and privately insured utilization in region s, let r^P_s denote the mean privately insured fee in region s, and let $e^z_s, z \in \{M, P\}$ denote the mean Medicare and privately insured expenditure in region s. Finally, let $\overline{\mu_s}$ denote the mean bargaining leverage in s and $\overline{\alpha_s}$ denote the mean capacity constraint in s.

Proposition SF1: There is significant across region-s variation in mean Medicare and privately insured spending that is not explained by variation in health status and outcomes i.e., $Var(e^M_s) > 0$ and $Var(e^P_s) > 0$ (see Appendix 3 for proof).

Across-region variations in Medicare mean spending ($Var(e^M_s)$) not coming from variations in health status must come from across-region variations in mean utilization ($Var(l^M_s)$),

as reimbursements rates are constant across regions (i.e., as $Var(F_s^M) = 0$). Across-region variations in mean Medicare utilization can come *directly* from across-region variations in mean capacity constraints, or equivalently, across-region variations in mean productivity (i.e., from $Var(\overline{Q}(\overline{\alpha_s}))$). They can also come *indirectly* from across-region variations in mean private reimbursement rates ($Var(r_s^P)$), or equivalently, from across region variation in mean bargaining leverage (i.e., $Var(\overline{\mu_s})$). Thus, elements which affect the private insurance market can spill-over to the Medicare market.

Across-region variations in privately-insured mean spending ($Var(e_s^P)$) not coming from variations in health status can come from across-region variations in mean utilization ($Var(l_s^P)$) and/or variations in mean private fees ($Var(r_s^P)$). Across-region variations in mean private utilization ($Var(l_s^P)$) can come from across-region variations in mean capacity constraints which are driven by across region variations in α (i.e., $Var(\overline{\alpha_s})$) as well as, across-region variation in mean private fees ($Var(r_s^P)$) which are driven by across region variations in μ (i.e., $Var(\overline{\mu_s})$). Thus, across-regions variations in Medicare and privately insured spending not driven by health status can be driven by variations in the models primitives α and μ.[19]

Proposition SF2: Across-region variation in mean Medicare spending ($Var(e_s^M)$) is driven principally by across-region variation in utilization ($Var(l_s^M)$), while privately insured spending variation ($Var(e_s^P)$) is primarily driven by across-region variation in the private fee ($Var(r_s^P)$) (see Appendix 4 for proof).

The model generates across-region variation in mean Medicare spending driven entirely by across-region variation in utilization (as Medicare reimbursement rates are constant across physicians and regions). As pointed out above in Proposition SF1, this across-region variation in Medicare utilization can be driven *directly* by across-region variations in mean capacity constraints (which are a function of the mean productivity parameter α) and *indirectly* by across-region variation in mean private fees (which is a function of the mean bargaining

[19]It is important to specify that for the aforementioned across-region variations in Medicare and privately insured spending *not* to lead to corresponding variations in outcomes, physicians must be practicing on the relatively flat portion of the health production function. We return to this point in Proposition SF7.

leverage parameter μ).

The model generates across-region variation in mean private spending which is driven primarily by across region variation in mean private fees, and not variations in quantities, if greater bargaining leverage translates into greater private fees while cost controls limit the crowding out of care to Medicare patients. The same result will also hold if bargaining leverage (μ) and productivity (α) are sufficiently negatively correlated across regions such that the price effect cancels out the capacity constraint effect.

Proposition SF3: Private provider prices are positively correlated with private spending ($Corr(F_s^P, e_s^P) > 0$) and negatively correlated with Medicare spending ($Corr(F_s^P, e_s^M) < 0$)

A sufficient condition for the across-region correlation between private prices and private spending to be positive (i.e., $Corr(F_s^P, e_s^P > 0)$) is that private prices and private utilization are positively correlated across regions (i.e., $Corr(F_s^P, l_s^P) > 0$). Nonetheless, $Corr(F_s^P, e_s^P) > 0$ allows for private prices to be negatively correlated with private quantities ($Corr(F_s^P, l_s^P) \leq 0$) across regions as long as this negative correlation is not too strong. Let us consider each of these possibilities separately to consider their implication on the second part of the proposition (i.e., $Corr(F_s^P, e_s^P) < 0$).

First, consider the case where private prices are positively correlated with private quantities (i.e., $Corr(F_s^P, l_s^P) > 0$) (a sufficient condition noted in the above paragraph). For the second part of the proposition to hold (i.e., that private prices are negatively correlated with Medicare spending), it must be the case that private prices are negatively correlated with Medicare utilization (i.e., $Corr(F_s^P, l_s^M) < 0$), as Medicare prices are fixed. This, in turn, would imply that regions with higher private prices would have, on average, higher private utilization and lower Medicare utilization (which would imply a negative correlation between private and Medicare utilization (i.e., $Corr(l_s^P, l_s^M) < 0$)).

However, if private prices are negatively correlated with private utilization ($Corr(F_s^P, l_s^P) \leq 0$), yet weak enough to allow for a positive correlation between private prices and private spending, and, also negatively correlated with Medicare quantities (i.e., $Corr(F_s^P, l_s^M) < 0$)

(given that Medicare prices are constant across regions) as given in the proposition, then private and Medicare quantities will be positively correlated ($corr(l_s^P, l_s^M) > 0$). That is regions with higher private prices have both lower private and Medicare utilization. However, the only way for both Medicare and private quantities to be lower under binding capacity constraints is for these same regions to have lower capacity . Thus, private prices (or, equivalently, bargaining leverage) and capacity constraints (or, equivalently, productivity) must be negatively correlated across regions for the proposition to hold under a (although weak) negative across-region correlation between private prices and private quantities.

Proposition SF4: Medicare and privately insured utilization is positively *correlated* across regions i.e., $Corr(l_s^M, l_s^P) > 0$

According to the previous derivation, in order for SF3 (i.e., $Corr(F_s^P, e_s^P) > 0$ and $Corr(F_s^P, e_s^M) < 0$) *as well as* SF4 ($Corr(l_s^M, l_s^P) > 0$) to hold, it must be the case: (i) private prices are negatively correlated with private utilization (i.e., $Corr(F_s^P, l_s^P) < 0$), and (ii) that private prices and capacity constraint are negatively correlated across regions (i.e., $corr(\mu, \alpha) < 0$) (which informs the relationship provided by $\Gamma(\mu_s, \alpha_s; \omega_s)$).

Proposition SF5: Medicare and privately insured spending is positively but weakly *correlated* across geographies i.e., $Corr(e_s^M, e_s^P) > 0$. (see Appendix 7 for proof).

From proposition SF4, we know that private and Medicare utilization are positively correlated across regions (i.e., $Corr(l_s^M, l_s^P) > 0$). We also know that the correlation between private Medicare spending is driven by the correlation between private spending and Medicare utilization as Medicare prices are constant across regions. The correlation between private spending and Medicare spending will be weaker than private utilization and Medicare utilization (i.e., $Corr(e_s^P, e_s^M) < Corr(l_s^P, l_s^M)$), if private prices and private utilization are negatively correlated across regions (i.e., $Corr(F_s^P, l_s^P) < 0$) which was established in Proposition SF3.

Proposition SF6: Increases in Medicare payments *cause* privately insured reimbursements to rise i.e., $\frac{\partial F_s^P}{\partial F^M} > 0$ (see Appendix 8 for proof).

Unlike many of the previous propositions examined across-region variations, this proposition considers within-region ones. More specifically, it states that an increase in the Medicare fee leads to an increase in the private fee within the same region. In order to examine the effect of an increase in the Medicare fee (F^M) on the commercial fee (F^P), we return to the endogenous price setting section. Notice that the increase in the F^M has two separate effects. First, it increases the disagreement utility by increasing the expected utility associated with turning away the current private patient. Secondly, conditional on the agreement utility being greater than the disagreement utility, it crowds-out the provision of care to the private patient. Thus, a private insurance provider which does not respond to an increase in the Medicare reimbursement rate risks having its patient receive less care from their current physician at best, and being dropped by their current physician all together at worse. In order to maximize the insurance provider's profit/utility, it must respond with an appropriate increase in its fee (F^P).

Proposition SF7: Variations in utilization (and spending) do not translate into corresponding variations in health outcome.

[*** Limit to across region at the mean?***] Our model is consistent with Proposition SF7 as long as the within and across equilibria Medicare and private utilization are on the relatively flat part of the health production function. More specifically, consider a set of physicians in region s, each treating a Medicare and privately insured patient with illness severity θ^M and θ^P. Given that each physician j within s is characterized by a unique bargaining leverage μ_j and productivity α_j, each will be associated with a unique pair of utility maximizing quantity of care $q^{*M}_{j,i,s}(\theta^M, \theta^P)$ and $q^{*P}_{j,k,s}(\theta^M, \theta^P)$, and by extension, a unique pair of patient health outcomes $h(\theta^M_i, q^{*M}_{j,i,s})$ and $h(\theta^P_k, q^{*P}_{j,k,s})$. Denote the within-region-s distribution of equilibrium health outcomes for Medicare and privately insured (where all Medicare patients suffer from the same illness θ^M and where all privately insured patients suffer from the same illness θ^P), patients as $S^M(q^{*M}_{j,i,s})$ and $S^P(q^{*P}_{j,k,s})$, respectively. Proposition SF7 simply requires that $h(\theta^M_i, q^{*M}_{j,i,s})$ is relatively constant across draws from $S^M(q^{*M}_{j,i,s})$ and

similarly that $h(\theta_k^P, q_{j,k,s}^{*P})$ is relatively constant across draws from $S^P(q_{j,k,s}^{*P})$. Or, equivalently, that $\frac{\partial h}{\partial q} \simeq 0$ along the distribution of $S^M(q_{j,i,s}^{*M})$ and $S^P(q_{j,k,s}^{*P})$.

4.1 Policy implications

In this section we introduce a series of changes to the environment. More specific, we derive the model's implication with respect to several policy changes. More specifically, we examine the likely implications of (i) an increase in provider bargaining power (or, equivalently, reduced competition in the provider market), (ii) the introduction of a pay-for-performance (P4P) bonus scheme to the Medicare market, (iii) the introduction of a capitation payment model to the private market, and, finally, (iv) an increase in the cost controls in the private market.

Theorem 1: Private utilization and spending increase while Medicare utilization and spending decrease, with an increase in provider bargaining power.

Proof:

Consider the impact of an increase in the physician's bargaining power μ_j on the equilibrium provision of private care:

$$\frac{dq_{j,k}^P}{d\mu_j} = -\frac{\frac{\partial R}{\partial \mu_j}}{\frac{\partial R}{\partial q_{j,k}^P}},$$

where

$$\frac{\partial R}{\partial \mu_j} = -F_{j,k,s}^{\prime P}(\mu_j) + f_{q^P,\mu}^{\prime\prime}(F_k^P(\mu_j)q_k^P) < 0,$$

given that $F_{j,k}^{\prime P}(\mu_j) > 0$ and $f_{q^P,\mu}^{\prime\prime}(F_{j,k}^P(\mu_j)q_{j,k}^P) < 0$. Furthermore, we know that:

$$\frac{\partial R}{\partial q_{j,k}^P} = -h_{22}^{\prime\prime}(\theta^M, \overline{Q}(\alpha) - q_{j,k}^P) - h_{22}^{\prime\prime}(\theta^P, q_{j,k}^P) + f_{q^P}^{\prime\prime}(F_{j,k}^P(\mu_j)q_{j,k}^P) > 0,$$

from Appendix 1 and the assumptions therein. Taken together,

$$\frac{dq^P_{j,k}}{d\mu_j} = \frac{F'^P_{j,k}(\mu_j) - f''_{12}(q^P_{j,k}, \mu_j)}{-h''_{22}(\theta^M, \overline{Q}(\alpha) - q^P_{j,k}) - h''_{22}(\theta^P, q^P_{j,k}) + f''_{q^P}(F^P_{j,k}(\mu_j)q^P_{j,k})} > 0.$$

Thus, an increase in the physician's bargaining power leads to an increase in private fee and a decrease in cost controls, which in turn lead to an increase in private utilization and spending . Furthermore, the increase in private utilization crowds-out Medicare utilization. As Medicare fees are invariant to physician bargaining power, a decrease in Medicare utilization is associated with a decrease in Medicare spending.

Similarly, a decrease in the physician's bargaining power will lead to a decrease in the private fee and an increase in cost controls, which in turn lead to a decrease in private utilization and spending. The decrease in private utilization leads to an increase in Medicare utilization. As Medicare fees are invariant to physician bargaining power, an increase in Medicare utilization is associated with an increase in Medicare spending.

Theorem 2: Under sufficiently large bonus payments, the introduction of a Pay-for-Performance scheme in the Medicare system will lead to an increase in Medicare utilization and spending, and a corresponding decrease in private utilization and spending.

Proof:

Consider a P4P system which rewards physicians for meeting quantity targets. More specifically consider that for each illness severity θ^M the Medicare authority provides an additional payment B if the care provided $q^M(\theta^M)$ is greater than some pre-determine quantity target $\overline{q}^M(\theta^M)$ (which we assume is costlessly verifiable). The physician's objective function thus becomes:

$$max\mathcal{L}_{q^M,q^P,\lambda} = h(\theta^M_i, q^M_i) + F^M q^M_i + B\mathbb{1}[q^M_i > \overline{q}(\theta)] + h(\theta^P_k, q^P_k)$$
$$+F^P_{j,k}(\mu_j)q^P_k - c(q^M_i + q^P_k) - f(F^P_{j,k}(\mu_j)q^P_k) + \lambda(\overline{Q}(\alpha_j) - q^M_i - q^P_k),$$

where $\mathbb{1}[q^M_i > \overline{q}(\theta^M)]$ is an indicator function which takes the value of 1 when the Medicare

quantity equals or exceeds the illness-specific quantity target ($\overline{q}(\theta^M)$). In order to solve the above program, the physician simply has to compare the maximized utility under the previous set-up to the utility with the targeted illness-specific quantity $\overline{q}(\theta^M)$ (the physician would never want to exceed $\overline{q}(\theta^M)$ as the marginal benefit of doing so would be outweighed by the marginal cost).

Notice that by providing the target quantity $\overline{q}(\theta)$, the physician's utility is given by:

$$\begin{aligned} U^{P4P}(\overline{q}^M(\theta), \overline{Q} - \overline{q}^M; B) = h(\theta_i^M, \overline{q}^M) + F^M \overline{q}^M + B + h(\theta_k^P, \overline{Q}(\alpha_j) - \overline{q}^M) \\ + F_{j,k}^P(\mu_j)(\overline{Q}(\alpha_j) - \overline{q}^M) - c(\overline{Q}(\alpha_j)) - f(\overline{Q}(\alpha_j) - \overline{q}^M, \mu_j) \end{aligned} \tag{13}$$

We know that for $B = 0$, $U^{P4P}(\overline{q}^M, \overline{Q} - \overline{q}^M; B = 0) < U(q^{*M}, q^{*P})$. As a result, there exists a $\overline{B}$ such that the LHS=RHS. That is, there exists a unique P4P bonus to induce physicians to provide a target level of care to Medicare patients. Offering a larger P4P bonus (i.e., beyond $\overline{B}$) would not further incentivize the physician to provide additional care to Medicare patients (and would simply result in a greater transfer of income to the provider).[20]

Theorem 3: The introduction of a capitation payment schemes for private payers leads to a decrease in private utilization (and potentially, private spending) and a corresponding increase in Medicare utilization and spending.

Proof: Reconsider the physician's objective function in the presence of a capitation payment K:

$$\begin{aligned} max \mathcal{L}_{q^M, q^P, \lambda} = h(\theta_i^M, q_i^M) + F^M q_i^M + h(\theta_k^P, q_k^P) \\ + K_{j,k}^P - c(q_i^M + q_k^P) + \lambda(\overline{Q}(\alpha_j) - q_i^M - q_k^P), \end{aligned}$$

where cost-controls are omitted as they are no longer relevant in the presence of a prospective payment system.

[20] This, of course, could lead to a response from the commercial insurance provider which could dampen the effect of P4P incentives in the Medicare market.

The new FOC conditions with respect to q^M, q^P and λ are given by:

$$h_2'(\theta_i^M, q_i^M) + F^M - c'(q_i^M + q_k^P) - \lambda = 0, \tag{14}$$

$$h_2'(\theta_k^P, q_k^P) - c'(q_i^M + q_k^P) - \lambda = 0, \tag{15}$$

$$\overline{Q}(\alpha_j) - q_i^M - q_k^P = 0. \tag{16}$$

The first-two FOCs can be simplified to:

$$h_2'(\theta_i^M, q_i^M) + F^M = h_2'(\theta_k^P, q_k^P). \tag{17}$$

Finally, by substituting the capacity constraint (which we assume is binding) into this last equality, the optimal amount of q^P must satisfy:

$$h_2'(\theta_i^M, \overline{Q}(\alpha_j) - q_k^P) + F^M = h_2'(\theta_k^P, q_k^P). \tag{18}$$

Introducing a capitation payment system will lead to a reduction in the equilibrium private utilization compared to the FFS system. The intuition is quite simple. When the physician provides a unit of care to the Medicare patient, she receives utility through two channels: the patient's improved health *and* the Medicare fee F^M. However, when providing a unit of care to the private patient, she receives utility uniquely through the improvement in the private patient's health. Whether or not reductions in private utilization leads to a reduction in private spending depends on the actual capitation payment K. Presumably, lower quantity provision by physicians paid by capitation (relative to their FFS counterparts) would translate into lower total costs.

Theorem 4: An increase in cost controls will lead to a decrease in private utilization

(with corresponding decrease in private spending) and an increase in Medicare utilization (with corresponding increase in Medicare spending).

Proof:

Consider an exogenous increase in cost control where $\overline{f}(F_q^P(\mu_j)q_k^P) > f(F_q^P(\mu_j)q_k^P)$ and $\overline{f}'(F_q^P(\mu_j)q_k^P) > f'(F_q^P(\mu_j)q_k^P)$ for each value of q^P and μ_j. The optimal amount of q^P must now satisfy $h_2'^M(\theta_i^M, \overline{Q}(\alpha_j) - q_k^P) + F^M = h_2'^P(\theta_k^P, q_k^P) + F_{j,k}^P(\mu_j) - \overline{f}'_{q^P}(F_q^P(\mu_j)q_k^P)$. Notice that the RHS (which is the net marginal benefit of providing q^P to the private patient) is smaller at equilibrium when compared to the previous level of monitoring. Thus, for the LHS to equate with this new level of net benefit to the private patient, the physician must provide more quantity to the Medicare patient in order to drive down the marginal benefit of care to this new level. This increase in Medicare utilization crowds out private utilization through the binding capacity constraint. Given that prices are invariant to such changes, the increase Medicare utilization is also associated with higher Medicare spending while the lower private utilization is associated with lower private spending.

5 Calibration

5.1 The maximization problem and optimal solution

We assume that the physician has the following maximization problem:

$$\max_{q^M, q^P} U = \beta \ln\left(q^M - \theta\right) + F^M q^M + \beta \ln\left(q^P - \theta\right) + \mu_j F^M q^P - \gamma \mu_j F^M q^P - \delta\left(q^M + q^P\right)^2$$
$$s.t. q^P + q^M = \alpha_j \overline{Q}$$

Plugging in the constraints and the maximization becomes unconstrained:

$$\max_{q^M, q^P} U = \beta \ln\left(\alpha_j \overline{Q} - q^P - \theta\right) + F^M\left(\alpha_j \overline{Q} - q^P\right) + \beta \ln\left(q^P - \theta\right) + (1-\gamma)\,\mu_j F^M q^P - \delta \alpha_j \overline{Q}^2$$

The FOC is given by:

$$\frac{1}{q^P-\theta}-\frac{1}{\alpha_j\overline{Q}-q^P-\theta}=\frac{F^M}{\beta}\left[1-(1-\gamma)\,\mu_j\right]\equiv C_j \tag{19}$$

which can be rewritten as:

$$\frac{\partial U}{\partial q^P}=\frac{C_j{q^P}^2-\left(\alpha_jC_j\overline{Q}+2\right)q^P+\alpha_j\overline{Q}+\alpha_jC_j\overline{Q}\theta-\theta^2C_j}{\left(\alpha_j\overline{Q}-q^P-\theta\right)\left(q^P-\theta\right)}=0 \tag{20}$$

The sign of C_j is critical to solution of q^P. From here on, we assume that $C_j>0$, which in turn implies the smaller solution of the numerator in (20) is delivering a local maximum. Meanwhile, by assuming $C_j>0$, we derive $q^P<q^M$.

Going back, the above FOC can be reorganized as below:

$${q^P}^2-\left(\alpha_j\overline{Q}+\frac{2}{C_j}\right)q^P+\frac{\alpha_j\overline{Q}}{C_j}+\alpha_j\overline{Q}\theta-\theta^2=0$$

There are two solutions to the above FOC, we can show that the smaller one achieves maximum, where $q^P=\frac{\alpha_j\overline{Q}C_j+2-\sqrt{B}}{2C_j}$ where $B=\left(\alpha_j\overline{Q}C_j\right)^2+4+4C_j^2\left(\theta^2-\alpha_j\overline{Q}\theta\right)$. Therefore, we find the optimal q^P and q^M as

$$\begin{aligned}q^P&=\tfrac{\alpha_j\overline{Q}}{2}+\tfrac{1}{C_j}-\sqrt{\tfrac{\left(\alpha_j\overline{Q}\right)^2}{4}+\tfrac{1}{C_j^2}+\left(\theta^2-\alpha_j\overline{Q}\theta\right)}\\ q^M&=\tfrac{\alpha_j\overline{Q}}{2}-\tfrac{1}{C_j}+\sqrt{\tfrac{\left(\alpha_j\overline{Q}\right)^2}{4}+\tfrac{1}{C_j^2}+\left(\theta^2-\alpha_j\overline{Q}\theta\right)}\end{aligned} \tag{21}$$

$q^M>q^P$ as expected.

5.2 Empirics data

The objective of this section is to derive moments to calibrate the model. To start, we used the following proxies to match model variables: q^P: private outpatient visits; q^M: Medicare outpatient visits. To derive variable μ, the following procedures is followed:

1. Let s_j^P be private outpatient spending for HRR j. We derive $f_j^P=\frac{s_j^P}{q_j^P}$ as the outpatient

fees.

2. We then find the average fee of private outpatient visit, weighted by the number of observations in each HRR, $\overline{f^P}$. (In the dataset we used, Acumen Medicare Aggregate HRR and LewinaggregateHRR, the mean is very close to median).

3. The medicare fee is then calculated as $f^M = 0.8\overline{f^P}$.

4. Then $\mu_j = \frac{f_j^P}{f^M}$.

We add the proxy of α using normalized sum of q^M and q^P. Namely let $\overline{Q} = \frac{\sum\left(q_i^P + q_i^M\right)}{N}$ and then $\alpha_i = \frac{q_i^P + q_i^M}{\overline{Q}}$. One most desirable property of α_i is that $corr\left(\alpha_i, \mu_i\right) < 0$ .In total our model produces four outcome variables $\left\{\alpha_i, \mu_i, q_i^P, q_i^M\right\}$. We then calculate the mean of the variable list $\hat{m}$, the standard error $\hat{\Sigma}$ and the correlation matrix $\hat{\varrho}$. Given the targetted moments m, Σ, ϱ as below:

$$m = [1, 1.25, 7.01, 2.82]$$

$$\Sigma = [0.11, 0.26, 2.24, 1.95]$$

$$\varrho = \begin{bmatrix} 1 & & & \\ -0.36 & 1 & & \\ 0.95 & -0.26 & 1 & \\ 0.77 & -0.42 & 0.53 & 1 \end{bmatrix}$$

5.3 Calibration

We propose a method that incorporates correlated α, μ and imposes little assumptions on the distribution of α, μ. To do so, we assign a chi-square distribution to the following terms $A_i = \frac{\left(\alpha_i \overline{Q}\right)}{2} - \theta, B_i = \frac{1}{C_i}$, where $\frac{1}{\beta}\left[F^M - (1-\gamma) F^M \mu_i\right] \equiv C_i$. And the optimal q^P, q^M are as follows:

$$q_i^P = \frac{\alpha_i \overline{Q}}{2} + \frac{1}{C_i} - \sqrt{\frac{\left(\alpha_i \overline{Q}\right)^2}{4} + \frac{1}{C_i^2} + \left(\theta^2 - \alpha_i \overline{Q}\theta\right)}$$

$$q_i^M = \frac{\alpha_i \overline{Q}}{2} - \frac{1}{C_i} + \sqrt{\frac{\left(\alpha_i \overline{Q}\right)^2}{4} + \frac{1}{C_i^2} + \left(\theta^2 - \alpha_i \overline{Q}\theta\right)}$$

or

$$q_i^P = A_i + B_i + \theta - \sqrt{A_i^2 + B_i^2}$$
$$q_i^M = A_i - B_i + \theta + \sqrt{A_i^2 + B_i^2}$$

Now we assume that $A_i = a_1 X_i + a_2 Y_i, B_i = b_1 X_i + b_2 Y_i$. We further assume that X_i, Y_i are mutually independent. To help us remove the square root in q^P, q^M , we impose the restriction $a_1 b_2 = a_2 b_1$ and we find that $\sqrt{A_i^2 + B_i^2} = c_1 X_i + c_2 Y_i$ where $c_1 = \sqrt{a_1^2 + a_2^2}, c_2 = \sqrt{b_1^2 + b_2^2}$. Thus q^P, q^M can now be rewritten as:

$$q_i^P = (a_1 + b_1 - c_1) X_i + (a_2 + b_2 - c_2) Y_i + \theta$$
$$q_i^M = (a_1 - b_1 + c_1) X_i + (a_2 - b_2 + c_2) Y_i + \theta$$

Let $EX_i = E_X, EY_i = E_Y, VX_i = V_X, VY_i = V_Y$. Thus we have:

$$Eq_i^P = (a_1 + b_1 - c_1) E_X + (a_2 + b_2 - c_2) E_Y + \theta$$
$$Eq_i^M = (a_1 - b_1 + c_1) E_X + (a_2 - b_2 + c_2) E_Y + \theta$$
$$Vq_i^P = (a_1 + b_1 - c_1)^2 V_X + (a_2 + b_2 - c_2)^2 V_Y$$
$$Vq_i^M = (a_1 - b_1 + c_1)^2 V_X + (a_2 - b_2 + c_2)^2 V_Y$$
$$cov\left(q_i^P, q_i^M\right) = (a_1 + b_1 - c_1)(a_1 - b_1 + c_1) V_X + (a_2 + b_2 - c_2)(a_2 - b_2 + c_2) V_Y$$

Meanwhile, we also know that:

$$E\alpha_i = \frac{2a_1 E_X + 2a_2 E_Y}{\overline{Q}}$$
$$V\alpha_i = \frac{4a_1^2 V_X + 4a_2^2 V_Y}{\overline{Q}^2}$$
$$cov\left(\alpha_i, q_i^P\right) = \frac{2a_1(a_1 + b_1 - c_1) V_X + 2a_2(a_2 + b_2 - c_2) V_Y}{\overline{Q}}$$
$$cov\left(\alpha_i, q_i^M\right) = \frac{2a_1(a_1 - b_1 + c_1) V_X + 2a_2(a_2 - b_2 + c_2) V_Y}{\overline{Q}}$$

The statistics with μ is somewhat non-trivial:

$$E\mu_i = \frac{1}{(1-\gamma)F^M} - E\left(\frac{\beta}{(1-\gamma)F^M(b_1X_i+b_2Y_i)}\right)$$
$$V\mu_i = V\left(\frac{\beta}{(1-\gamma)F^M(b_1X_i+b_2Y_i)}\right)$$
$$cov\left(\alpha_i, \mu_i\right) = -cov\left(\frac{\beta}{(1-\gamma)F^M(b_1X_i+b_2Y_i)}, \frac{2a_1X_i+2a_2Y_i}{\overline{Q}}\right)$$
$$cov\left(\mu_i, q_i^P\right) = -cov\left(\frac{\beta}{(1-\gamma)F^M(b_1X_i+b_2Y_i)}, \left(a_1+b_1-c_1\right)X_i + \left(a_2+b_2-c_2\right)Y_i\right)$$
$$cov\left(\mu_i, q_i^M\right) = -cov\left(\frac{\beta}{(1-\gamma)F^M(b_1X_i+b_2Y_i)}, \left(a_1-b_1+c_1\right)X_i + \left(a_2-b_2+c_2\right)Y_i\right)$$

The calibration of these statistics require more specific assumption of X, Y (The uniform distributions might make things easy). Below we sum up the calibration scheme. We have 13 parameters in the model $\left\{a_1, a_2, b_1, b_2, E_X, E_Y, V_X, V_Y, \overline{Q}, \theta, F^M, \beta, \gamma\right\}$ to match 15 moments. In all, our task of calibration is reduced into solving the below system of nonlinear equations:

$$Eq_i^P = \left(a_1+b_1-c_1\right)E_X + \left(a_2+b_2-c_2\right)E_Y + \theta \tag{22}$$

$$Eq_i^M = \left(a_1-b_1+c_1\right)E_X + \left(a_2-b_2+c_2\right)E_Y + \theta \tag{23}$$

$$Vq_i^P = \left(a_1+b_1-c_1\right)^2 V_X + \left(a_2+b_2-c_2\right)^2 V_Y \tag{24}$$

$$Vq_i^M = \left(a_1-b_1+c_1\right)^2 V_X + \left(a_2-b_2+c_2\right)^2 V_Y \tag{25}$$

$$cov\left(q_i^P, q_i^M\right) = \left(a_1+b_1-c_1\right)\left(a_1-b_1+c_1\right)V_X + \left(a_2+b_2-c_2\right)\left(a_2-b_2+c_2\right)V_Y \tag{26}$$

$$E\alpha_i = \frac{2a_1E_X + 2a_2E_Y}{\overline{Q}} \tag{27}$$

$$V\alpha_i = \frac{4a_1^2 V_X + 4a_2^2 V_Y}{\overline{Q}^2} \tag{28}$$

$$cov\left(\alpha_i, q_i^P\right) = \frac{2a_1\left(a_1 + b_1 - c_1\right)V_X + 2a_2\left(a_2 + b_2 - c_2\right)V_Y}{\overline{Q}} \tag{29}$$

$$cov\left(\alpha_i, q_i^M\right) = \frac{2a_1\left(a_1 - b_1 + c_1\right)V_X + 2a_2\left(a_2 - b_2 + c_2\right)V_Y}{\overline{Q}} \tag{30}$$

$$E\mu_i = \frac{1}{\left(1-\gamma\right)F^M} - E\left(\frac{\beta}{\left(1-\gamma\right)F^M\left(b_1X_i + b_2Y_i\right)}\right) \tag{31}$$

$$V\mu_i = V\left(\frac{\beta}{\left(1-\gamma\right)F^M\left(b_1X_i + b_2Y_i\right)}\right) \tag{32}$$

$$cov\left(\alpha_i, \mu_i\right) = -cov\left(\frac{\beta}{\left(1-\gamma\right)F^M\left(b_1X_i + b_2Y_i\right)}, \frac{2a_1X_i + 2a_2Y_i}{\overline{Q}}\right) \tag{33}$$

$$cov\left(\mu_i, q_i^P\right) = -cov\left(\frac{\beta}{\left(1-\gamma\right)F^M\left(b_1X_i + b_2Y_i\right)}, \left(a_1 + b_1 - c_1\right)X_i + \left(a_2 + b_2 - c_2\right)Y_i\right) \tag{34}$$

$$cov\left(\mu_i, q_i^M\right) = -cov\left(\frac{\beta}{\left(1-\gamma\right)F^M\left(b_1X_i + b_2Y_i\right)}, \left(a_1 - b_1 + c_1\right)X_i + \left(a_2 - b_2 + c_2\right)Y_i\right) \tag{35}$$

$$a_1b_2 = a_2b_1 \tag{36}$$

There're several things to note in the above equation system:

1. With (28) and (29), we can derive (30), which reduces the total number of equations from 15 to 14.

2. Likewise, we can derive (35) from linear combination of equation (33) and (34). We are dropping equation (35) and now we have 13 equations remaining.

3. (30) has a close form:

$$cov\left(\alpha_i, \mu_i\right) = -cov\left(\frac{\beta}{(1-\gamma)F^M(b_1X_i+b_2Y_i)}, \frac{2a_1X_i+2a_2Y_i}{\overline{Q}}\right) = E\left(\frac{\beta(2a_1X_i+2a_2Y_i)}{(1-\gamma)F^M\overline{Q}(b_1X_i+b_2Y_i)}\right) - E\left(\frac{\beta}{(1-\gamma)F^M(b_1X_i+b_2Y_i)}\right)$$

$$cov\left(\alpha_i, \mu_i\right) = -cov\left(\frac{\beta}{(1-\gamma)F^M(b_1X_i+b_2Y_i)}, \frac{2a_1X_i+2a_2Y_i}{\overline{Q}}\right)$$

$$= \frac{\beta a_2}{(1-\gamma)F^M\overline{Q}b_2} - E\left(\frac{\beta}{(1-\gamma)F^M(b_1X_i+b_2Y_i)}\right)E\left(\frac{2a_1X_i+2a_2Y_i}{\overline{Q}}\right)$$

Now we are matching 13 moments with 13 parameters. It remains challenging to get the close form of equation (34). The key challenge is removing the covariance operator in the equation.

Moving on, we assume that $X_1 \sim Gamma\left(k_1, \delta_1\right), Y_1 \sim Gamma\left(k_2, \delta_2\right)$. We also assume that $b_1\delta_1 = b_2\delta_2$. This assumption is important since it allows that linear combination of X_1, Y_1 are also Gamma distributed. Since we add one more constraints, one of the above equations should be dropped. Our choice of the equation is equation 16 due to the difficulty in deriving a close form. Now we still have 13 parameters $\left\{a_1, a_2, b_1, b_2, k_1, k_2, \delta_1, \delta_2, \overline{Q}, \theta, F^M, \beta, \gamma\right\}$

.

$$Eq_i^P = \left(a_1 + b_1 - c_1\right)k_1\delta_1 + \left(a_2 + b_2 - c_2\right)k_2\delta_2 + \theta \tag{37}$$

$$Eq_i^M = \left(a_1 - b_1 + c_1\right)k_1\delta_1 + \left(a_2 - b_2 + c_2\right)k_2\delta_2 + \theta \tag{38}$$

$$Vq_i^P = \left(a_1 + b_1 - c_1\right)^2 k_1\delta_1^2 + \left(a_2 + b_2 - c_2\right)^2 k_2\delta_2^2 \tag{39}$$

$$Vq_i^M = \left(a_1 - b_1 + c_1\right)^2 k_1\delta_1^2 + \left(a_2 - b_2 + c_2\right)^2 k_2\delta_2^2 \tag{40}$$

$$cov\left(q_i^P, q_i^M\right) = \left(a_1 + b_1 - c_1\right)\left(a_1 - b_1 + c_1\right)k_1\delta_1^2 + \left(a_2 + b_2 - c_2\right)\left(a_2 - b_2 + c_2\right)k_2\delta_2^2 \quad (41)$$

$$E\alpha_i = \frac{2a_1k_1\delta_1 + 2a_2k_2\delta_2}{\overline{Q}} \quad (42)$$

$$V\alpha_i = \frac{4a_1^2k_1\delta_1^2 + 4a_2^2k_2\delta_2^2}{\overline{Q}^2} \quad (43)$$

$$cov\left(\alpha_i, q_i^P\right) = \frac{2a_1\left(a_1 + b_1 - c_1\right)k_1\delta_1^2 + 2a_2\left(a_2 + b_2 - c_2\right)k_2\delta_2^2}{\overline{Q}} \quad (44)$$

$$E\mu_i = \frac{1}{\left(1-\gamma\right)F^M} - \frac{\beta b_1\delta_1}{\left(1-\gamma\right)F^M\left(b_1k_1 + b_2k_2 - 1\right)} \quad (45)$$

$$V\mu_i = \frac{\left(\beta b_1\delta_1\right)^2}{\left[\left(1-\gamma\right)F^M\left(b_1k_1 + b_2k_2 - 1\right)\right]^2\left(b_1k_1 + b_2k_2 - 2\right)} \quad (46)$$

$$cov\left(\alpha_i, \mu_i\right) = \frac{\beta a_2}{\left(1-\gamma\right)F^M\overline{Q}b_2} - \frac{\beta b_1\delta_1\left(2a_1k_1\delta_1 + 2a_2k_2\delta_2\right)}{\left(1-\gamma\right)F^M\overline{Q}\left(b_1k_1 + b_2k_2 - 1\right)} \quad (47)$$

$$a_1b_2 = a_2b_1 \quad (48)$$

$$b_1\delta_1 = b_2\delta_2 \quad (49)$$

Equation 27-29 is derived as follows: $Z_1 = b_1X_i + b_2Y_i \sim Gamma\left(b_1k_1 + b_2k_2, b_1\delta_1\right)$ and thus we know that $\frac{1}{Z_1} \sim inv - Gamma\left(b_1k_1 + b_2k_2, b_1\delta_1\right)$.

[Results and Counterfactual Experiments]

6 Conclusions

In this paper we build a model which replicates the main stylized facts surrounding the geographic-variations' in healthcare utilization and spending literature. Not only does it replicate the within- and across-geographies patterns of utilization and spending in the Medicare and private settings, it replicates their interdependencies (in terms of fees, utilization and spending). These patterns are driven by the two primitives of the model: the physician's bargaining leverage and productivity.

Our model allows us to speak to the likely effects of different policies. We show that decreases in provider bargaining leverage (or, equivalently, increases in provider competition) negatively affect private reimbursement rates and cost controls, which in turn negatively affect private utilization and spending (with corresponding positive effects on Medicare utilization and spending). We also show that an increase in the Medicare reimbursement rate can have a causal effect on private reimbursement rates, utilization and spending in both the Medicare and private-insurance markets. Finally, we show that pay-for-performance in Medicare markets and prospective payments cost-controls in the private market can have important implications for both these markets - again highlighting their interdependencies. Taken together these results suggest that private and Medicare variations in utilization and spendings across geographies are driven, at least in part, by local market conditions and the incentives they create rather than just inherent differences in physician practice styles and patient need.

References

Chandra, A., Finkelstein, A., Sacarny, A., and Syverson, C. (2013). Healthcare exceptionalism? productivity and allocation in the us healthcare sector. Technical report, National Bureau of Economic Research.

Chandra, A. and Skinner, J. (2012). Technology growth and expenditure growth in health

care. *Journal of Economic Literature*, 50(3):645–680.

Chandra, A. and Staiger, D. O. (2007). Productivity spillovers in healthcare: evidence from the treatment of heart attacks. *The Journal of Political Economy*, 115:103.

Chernew, M. E., Sabik, L. M., Chandra, A., Gibson, T. B., and Newhouse, J. P. (2010). Geographic correlation between large-firm commercial spending and medicare spending. *The American Journal of Managed Care*, 16(2):131.

Clemens, J. and Gottlieb, J. D. (2017). In the shadow of a giant: Medicare's influence on private physician payments. *Journal of Political Economy*, 125(1):1–39.

Cutler, D., Skinner, J., Stern, A. D., and Wennberg, D. (2013). Physician beliefs and patient preferences: A new look at regional variation in health care spending. Working Paper 19320, National Bureau of Economic Research.

Dafny, L. S. (2005). How do hospitals respond to price changes? *American Economic Review*, pages 1525–1547.

Doyle, J. J., Graves, J. A., Gruber, J., and Kleiner, S. (2012). Do high-cost hospitals deliver better care? evidence from ambulance referral patterns. Technical report, National Bureau of Economic Research.

Dunn, A. and Shapiro, A. H. (2014). Do physicians possess market power? *Journal of Law and Economics*, 57(1):159–193.

Finkelstein, A., Gentzkow, M., and Williams, H. (2016). Sources of geographic variation in health care: Evidence from patient migration. *Quarterly Journal of Economics*, 131(4):1681–1726.

Finkelstein, A., Gentzkow, M., and Williams, H. (2018). Place-based drivers of mortality: Evidence from migration. *NBER Working Paper*, 15241.

Fisher, E. S., Bynum, J. P., and Skinner, J. S. (2009). Slowing the growth of health care costs?lessons from regional variation. *New England Journal of Medicine*, 360(9):849–852.

Fisher, E. S., Wennberg, D. E., Stukel, T. A., Gottlieb, D. J., Lucas, F. L., and Pinder, E. L. (2003a). The implications of regional variations in medicare spending. part 1: the content, quality, and accessibility of care. *Annals of internal medicine*, 138(4):273–287.

Fisher, E. S., Wennberg, D. E., Stukel, T. A., Gottlieb, D. J., Lucas, F. L., and Pinder, E. L. (2003b). The implications of regional variations in medicare spending. part 2: health outcomes and satisfaction with care. *Annals of internal medicine*, 138(4):288–298.

Frakt, A. B. (2011). How much do hospitals cost shift? a review of the evidence. *Milbank Quarterly*, 89(1):90–130.

Fuchs, V. R. (2004). More variation in use of care, more flat-of-the-curve medicine: Why does it occur? what should be done about it? *Health Affairs*, 23(Suppl2):VAR–104.

Gawande, A. (2009). The cost conundrum. *The New Yorker*, (June 1).

Gaynor, M., Ho, K., and Town, R. (2014). The industrial organization of health care markets. Technical report, National Bureau of Economic Research.

Glover, J. A. (1938). The incidence of tonsillectomy in school children. *Indian Journal of Pediatrics*, 5(4):252–258.

Gottlieb, D. J., Zhou, W., Song, Y., Andrews, K. G., Skinner, J. S., and Sutherland, J. M. (2010). Prices don't drive regional medicare spending variations. *Health Affairs*, pages 10–1377.

Hussey, P. S., Wertheimer, S., and Mehrotra, A. (2013). The association between health care quality and cost: a systematic review. *Annals of internal medicine*, 158(1):27–34.

Kleiner, S., White, W., and Lyons, S. (2015). Market power and provider consolidation in physician markets. *International Journal of Health Economics and Management*, 15(1):99–126.

McClellan, M. (1997). Hospital reimbursement incentives: an empirical analysis. *Journal of Economics & Management Strategy*, 6(1):91–128.

McGuire, T. and Pauly, M. (1991). Physician response to fee changes with multiple payers. *Journal of Health Economics*, 10(4):385–410.

McKellar, M. R., Naimer, S., Landrum, M. B., Gibson, T. B., Chandra, A., and Chernew, M. (2014). Insurer market structure and variation in commercial health care spending. *Health Services Research*, 49(3):878–892.

MedPAC (2011). Regional variation in medicare service use. In *Report to Congress*. MedPAC.

Newhouse, J. P., Garber, A. M., Graham, R. P., McCoy, M. A., Mancher, M., Kibria, A., et al. (2013). *Variation in Health Care Spending: Target Decision Making, Not Geography*. National Academies Press.

Philipson, T. J., Goldman, D. P., Seabury, S. A., Lakadawalla, D. N., and Lockwood, L. M. (2010). Geographic variation in health care: The role of private markets. *Brookings Papers on Economic Activity: Spring 2010*, page 325.

Romley, J. A., Axeen, S., Lakdawalla, D. N., Chernew, M. E., Bhattacharya, J., and Goldman, D. P. (2014). The relationship between commercial health care prices and medicare spending and utilization. *Health Services Research*.

Sheiner, L. (2014). Why the geographic variation in health care spending cannot tell us much about the efficiency or quality of our health care system. *Brookings Papers on Economic Activity*, 2014(2):1–72.

Silverman, E. and Skinner, J. (2004). Medicare upcoding and hospital ownership. *Journal of health economics*, 23(2):369–389.

Skinner, J. (2011). *Causes and consequences of regional variations in health care*, chapter 2, pages 49–93. Elsevier.

Skinner, J. and Staiger, D. (2007). Technological diffusion from hybrid corn to beta blockers. *Hard-to-Measure Goods and Services: Essays in Honor of Zvi Griliches. University of Chicago Press and NBER.*

Town, R., Feldman, R., and Kralewski, J. (2011). Market power and contract form: evidence from physician group practices. *International Journal of Health Care Finance and Economics*, 11(2):115–132.

Wennberg, J. and Gittelsohn, A. (1973). Small area variations in health care delivery a population-based health information system can guide planning and regulatory decision-making. *Science*, 182(4117):1102–1108.

Wennberg, J. E. and Cooper, M. M. (1996). *The Dartmouth atlas of health care*. American Hospital Publishing Chicago, Ill:.

Wennberg, J. E., Fisher, E. S., Skinner, J. S., et al. (2002). Geography and the debate over Medicare reform. *Health Affairs*, 21(2):10–10.

7 Appendices

Appendix 1:

First, consider the following additional set of conditions:

- $h''_{21}(\theta^M, q^M) > 0$ and $h''_{21}(\theta^P, q^P) > 0$ (the marginal benefit of care increases with illness severity), and
- $h''_{22}(\theta^M, q^M) < 0$ and $h''_{22}(\theta^P, q^P) < 0$ (the marginal benefit of care is decreasing in care).

Next, let R define the (implicit) supply function (8) such that:

$$R \equiv h'_2(\theta_i^M, (\overline{Q}(\alpha_j) - q_k^P)) + F^M$$
$$-h'_2(\theta_k^P, q_k^P) - F^P_{j,k,s}(\mu_j) + f'_{q^P}(F^P_{j,k,s}(\mu_j) q_k^P).$$

Using the implicit function theorem (and omitting subscripts for compactness), consider:

$$\frac{dq^P}{d\theta^P} = -\frac{\frac{\partial R}{\partial \theta^P}}{\frac{\partial R}{\partial q^P}}$$

where

$$\frac{\partial R}{\partial \theta^P} = -h''_{21}(\theta^P, q^P)$$

and where

$$\frac{\partial R}{\partial q^P} = -h''_{22}(\theta^M, \overline{Q}(\alpha) - q^P) - h''_{22}(\theta^P, q^P) + f''_{q^P}(F^P_{j,k,s}(\mu_j) q_k^P).$$

Thus,

$$\begin{aligned}\frac{dq^P}{d\theta^P} &= -\frac{-h''_{21}(\theta^P, q^P)}{-h''_{22}(\theta^M, \overline{Q}(\alpha) - q^P) - h''_{22}(\theta^P, q^P) + f''_{q^P}(F^P_{j,k,s}(\mu_j)q^P_k)} \\ &= -\frac{h''_{21}(\theta^P, q^P)}{h''_{22}(\theta^M, \overline{Q}(\alpha) - q^P) + h''_{22}(\theta^P, q^P) - f''_{q^P}(F^P_{j,k,s}(\mu_j)q^P_k)} \\ &> 0.\end{aligned}$$

By the capacity constraint, the above implies that

$$\frac{dq^M}{d\theta^P} < 0.$$

Similarly, we can show that under the same conditions:

$$\frac{dq^M}{d\theta^M} = -\frac{\frac{\partial R}{\partial \theta^M}}{\frac{\partial R}{\partial q^M}} > 0$$

and

$$\frac{dq^M}{d\theta^P} = -\frac{\frac{\partial R}{\partial \theta^P}}{\frac{\partial R}{\partial q^M}} < 0.$$

Letting $\Delta(\theta^M, \theta^P) \equiv q^P(\theta^M, \theta^P) - q^M(\theta^M, \theta^P)$:

$$\Delta'_2(\theta^M, \theta^P) = \frac{\partial q^{*P}}{\partial(\theta^P)} - \frac{\partial q^{*M}}{\partial(\theta^P)} > 0$$

and

$$\Delta'_1(\theta^M, \theta^P) = \frac{\partial q^{*P}}{\partial(\theta^M)} - \frac{\partial q^{*M}}{\partial(\theta^M)} < 0.$$

Appendix 2:

Consider the effect of a marginal increase in the capacity constraint ($\overline{Q}(\alpha)$) on the optimal provision of care to the private patient k (while dropping subscripts for compactness) while holding all other elements constant including the private fee (or equivalently, market power μ):

$$\frac{dq^P}{d\overline{Q}(\alpha)} = -\frac{\frac{\partial R}{\partial \overline{Q}(\alpha)}}{\frac{\partial R}{\partial q^P}}.$$

Given that:

$$\frac{\partial R}{\partial \overline{Q}(\alpha)} = h''_{22}(\theta^M, \overline{Q}(\alpha) - q^P),$$

$$\frac{dq^P}{d\overline{Q}(\alpha)} = -\frac{h''_{22}(\theta^M, \overline{Q}(\alpha) - q^P)}{-h''_{22}(\theta^M, \overline{Q}(\alpha) - q^P) - h''_{22}(\theta^P, q^P) + f''_{q^P}(F^P_{j,k,s}(\mu_j)q^P_k)}.$$

Under the conditions set out in Appendix 1,

$$0 < \frac{dq^P}{d\overline{Q}(\alpha)} < 1.$$

which in turn implies that:

$$0 < \frac{dq^M}{d\overline{Q}(\alpha)} < 1.$$

Further notice that:

$$0 < \frac{dq^P}{d\overline{Q}(\alpha)} < \frac{1}{2}.$$

if

$$-h''_{22}(\theta^M, \overline{Q}(\alpha) - q^P) < -h''_{22}(\theta^P, q^P) + f''_{q^P}(F^P_{j,k,s}(\mu_j)q^P_k), \tag{50}$$

which will hold for certain if $h''_{22}(\theta^M, \overline{Q}(\alpha) - q^P) = h''_{22}(\theta^P, q^P)$ given the assumption that $f''_{q^P}(F^P_{j,k,s}(\mu_j)q^P_k) > 0$.

This above also implies that:

$$\frac{dq^M}{d\overline{Q}(\alpha)} > \frac{1}{2}.$$

Appendix 3:

Consider a geographic location s with J_s physicians. Further assume that each physician j's capacity constraint and market power parameters are drawn from a known region-s specific distribution $\Gamma(\alpha, \mu; \omega_s)$, where ω_s is drawn from known distribution $K(\omega_s)$. Finally, assume that each physician treats two patients (one private and one Medicare) with illness severities θ^P and θ^M which are common across *all* physicians irrespective of region s (and thus do not drive the across-region variances).

Although each physician (within the same region s) faces the same medicare fee F^M, they face (i) different private fees ($F^P_{j,s}(\mu_j)$), (ii) different cost controls ($f(F^P_{j,k,s}(\mu_j)q^P_k)$), and (ii) different capacity constraints $\overline{Q}(\alpha_j)$. Thus, each physician j's utility maximizing supply of care to the Medicare patient and the private patient (again within the same region s) are given by: $q^{*M}_{j,s} = \mathcal{Q}(F^M, F^P_{j,s}, \theta^P, \theta^M; \alpha_j, \mu_j)$ and $q^{*P}_{j,s} = \mathcal{Q}(F^M, F^P_{j,s}, \theta^P, \theta^M; \alpha_j, \mu_j)$, respectively.

Thus, for a given pair of illness severities, the distribution $\Gamma(\alpha, \mu; \omega_s)$ yields a unique region-s specific distribution of private fees, capacity constraints, as well as private and Medicare quantities with corresponding region-s specific (i) mean private fee (r^P_s), (ii) mean private (l^P_s) and Medicare (l^M_s) utilization, and (iii) mean private (e^P_s) and Medicare (e^M_s) spending. Similarly, for the same pair of illness severities, the distribution $\Gamma(\alpha, \mu; \omega_{s'})$ (with $J_{s'}$ physicians) yields its own unique region s' distribution of fees as well as private and Medicare quantities with corresponding $r^P_{s'}$, $l^P_{s'}$, $l^M_{s'}$, $e^P_{s'}$ and $e^M_{s'}$.

Considering the entire set of S regions, (i) across-region mean Medicare and Private utilization are given by $E(l_s^M) = \frac{\sum_{s=1}^{S} l_s^M}{S}$ and $E(l_s^P) = \frac{\sum_{s=1}^{S} l_s^P}{S}$, (ii) aross-region mean private fees in given by: $E(r_s^P) = \frac{\sum_{s=1}^{S} r_s^P}{S}$, (iii) across-region mean Medicare and Private spending are given by $E(e_s^M) = \frac{\sum_{s=1}^{S} e_s^M}{S}$ and $E(e_s^P) = \frac{\sum_{s=1}^{S} e_s^P}{S}$ (all of which are expectations taken over the region-specific means). We denote: (i) the across-region variance of mean Medicare and private utilization as $Var(l_s^M)$ and $Var(l_s^P)$, (ii) the across-region variance of mean private fees as $Var(r_s^P)$, and (iii) the across-region variance of mean Medicare and private spending as $Var(e_s^M)$ and $Var(e_s^P)$.

Appendix 4:

Because Medicare fees are held constant across regions, the only endogenous source of across-region variation in mean Medicare spending is variation in across-region mean utilization. That is, $Var(e_s^M)$=$Var(F^M l_s^M)$=$(F^M)^2 Var(l_s^M)$. Thus, by construction, across region variation in mean Medicare spending *must* come from the across region variation in mean Medicare utilization i.e., $Var(l_s^M)$. The sources of this across region variation in mean Medicare utilization are described explicitly in the Proposition SF1.

With respect to the private market, across-region variation in mean spending can theoretically be driven by both across-region variations in mean private fees *as well as* across region variation in mean utilization where $Var(e_s^P) = Var(r_s^P l_s^P)$. The across region variation in mean spending $Var(e_s^P)$ will be driven by across-region variations in private fees $Var(r_s^P)$ and not across-region variation in mean utilization $Var(l_s^P)$ if there is considerable across region variation in the mean physician bargaining leverage ($\overline{\mu}$) which (i) leads to important across-region variations in the mean private fees (i.e., $F'(\overline{\mu}) > 0$), *but* (ii) also has little impact on private volume due to either very strong cost controls limiting physicians' ability to respond to higher private fees with greater volume, or, there exists a sufficiently negative across-region correlation between α and μ such that the positive price effect is counteracted by the negative capacity constraint effect.

To see this, consider the extreme case where the across-region variance is driven entirely

by the across-region variance in private fees or $Var(e_s^P) = Var(r_s^P l^P) = (l^P)^2 Var(r_S^P)$ and where $Var(r_s^P)$ is driven by across-region variation in mean bargaining $\overline{\mu}$ leverage.

Assume that mean capacity constraints are held constant across regions. For simplicity, also consider that each region is populated by one physician (a representative mean physician) and thus q and l, as well as, F and r, are interchangeable. A higher private mean fee F_s^P will not translate into higher or lower mean utilization across regions if:

$$\frac{dq_s^P}{dF_s^P} = -\frac{\frac{\partial R}{\partial F^P}}{\frac{\partial R}{\partial q^P}} = 0, \tag{51}$$

where:

$$\frac{\partial R}{\partial F_s^P(\mu)} = -1 + f''_{q^P,F_s}(F_s^P q_s^P)$$

and where,

$$\frac{\partial R}{\partial q_s^P} = -h''_{22}(\theta^M, \overline{Q}_s(\alpha) - q_s^P) - h''_{22}(\theta^P, q_s^P) + f''_{q_s^P}(F_s^P q_s^P).$$

By substitution:

$$\frac{dq^P}{dF_s^P(\mu)} = -\frac{-1 + f''_{F_s^P,q_s^P}(F_s^P q_s^P)}{-h''_{22}(\theta^M, \overline{Q}_s(\alpha) - q_s^P) - h''_{22}(\theta^P, q_s^P) + f''_{q_s^P}(F_s^P q_s^P)} = 0$$

if f''_{q^P,F^P}=1. That is, the price effect is completely countered by the cost control effect.

If, however, mean capacity constraints (i.e., mean α) are negatively correlated with mean fees (i.e., mean μ) across regions, then across region utilization will not vary if the capacity effect exactly cancels out the price effect. That is, $\frac{dq^P}{dF(\mu)} + \frac{dq^P}{dQ(\alpha)} = 0$ for a positive change in μ and corresponding negative change in α given by $\Gamma(\alpha, \mu; \omega_s)$ and $K(\omega)$.

Appendix 6:

Holding everything else constant (including market power and, consequently, private fees), the supply of Medicare *and* private quantities is increasing in the physician's capacity con-

straint $\overline{Q}(\alpha)$ which is itself increasing in the physician's productivity (α) (see Appendix 2). Thus, $corr(l_s^M, l_s^P) > 0$. If, however, an increase in capacity constraint is associated with a decrease in private private (i.e., if μ and α are negatively correlated across regions).

Appendix 7: For Proposition SF5 to hold in light of Proposition SF4, $0 < corr(e_s^M, e_s^P) < corr(l_s^M, l_s^P)$.

First consider,

$$corr(e_s^M, e_s^P) = \frac{cov(e_s^M, e_s^P)}{sd(e_s^M)sd(e_s^P)} \tag{52}$$

or

$$corr(e_s^M, e_s^P) = \frac{cov(r^M l_s^M, r_s^P l_s^P)}{sd(r^M l_s^M)sd(r_s^P l_s^P)} \tag{53}$$

Next consider,

$$corr(l_s^M, l_s^P) = \frac{cov(l_s^M, l_s^P)}{sd(l_s^M)sd(l_s^P)}. \tag{54}$$

Now, SF5 will hold if:

$$\frac{cov(r^M l_s^M, r_s^P l_s^P)}{sd(r^M l_s^M)sd(r_s^P l_s^P)} < \frac{cov(l_s^M, l_s^P)}{sd(l_s^M)sd(l_s^P)} \tag{55}$$

or

$$\frac{E(r^M l_s^M r_s^P l_s^P) - E(r^M l_s^M)E(r_s^P l_s^P)}{r^M sd(l_s^M)sd(r_s^P l_s^P)} < \frac{E(l_s^M l_s^P) - E(l_s^M)E(l_s^P)}{sd(l_s^M)sd(l_s^P)} \tag{56}$$

or

$$\frac{E(l_s^M r_s^P l_s^P) - E(l_s^M)E(r_s^P l_s^P)}{sd(l_s^M)sd(r_s^P l_s^P)} < \frac{E(l_s^M l_s^P) - E(l_s^M)E(l_s^P)}{sd(l_s^M)sd(l_s^P)} \tag{57}$$

as r^M is constant across regions.

Notice that the denominator of LHS is greater than the denominator of the RHS if $sd(r_s^P l_s^P) > sd(l_s^P)$. But we know this to be the case as the variation in private spending is driven primarily by across-region variations in mean prices rather than mean quantities (as spelled out in SF2).

Thus, a sufficient condition for SF5 to hold is that:

$$E(l_s^M r_s^P l_s^P) - E(l_s^M)E(r_s^P l_s^P) < E(l_s^M l_s^P) - E(l_s^M)E(l_s^P) \tag{58}$$

or, equivalently,

$$cov(l_s^M, r_s^P l_s^P) < cov(l_s^M, l_s^P) \tag{59}$$

.

But we know this to be the case as r_s^P and l_s^M are negatively correlated. That is, although quantities will be correlated positively across regions as they are both positively correlated with the region-specific capacity constraint, higher private fees are nonetheless negatively correlated with public provision/consumption.

Appendix 8:

Consider an increase in the Medicare reimbursement rate F^M. Such an increase will potentially have two distinct effects on the provision of care to the private patient k. First, as is shown in Proposition SF3 and its corresponding proof in Appendix 5, ceteris paribus, an increase in the Medicare fee leads to an increase in the provision of care to the Medicare patient and a corresponding crowding out of private quantities. Second, an increase in the Medicare fee leads to an increase in the physician's disagreement utility U^D.

Consider the first effect on the insurance provider's agreement utility. Let $V_k^{A,F^M} = V(h(\theta_k^P, q_k^{*P}), F_j^P q_k^{*P})$ be the insurer's equilibrium agreement utility under the initial Medicare fee F^M. Furthermore, let $V_k^{A,F'^M} = V(h(\theta_k^P, q_k^{'*P}), F_j^P q_k^{'*P})$ be the insurer's new agreement utility under new F^M - where $q_k^{*P} > q_k^{'*P}$. Notice that this decrease in the provision of medical care to the private patient k (due to the crowd-out) leads to a decrease in the insurance provider's utility through the patient's health but also an increase in the insurance provider's utility/profit due to the corresponding cost savings. However, given the move away from the equilibrium where the marginal cost of care was equal to its marginal benefit (from the insurance provider's perspective), the insurance provider's utility/profit is reduced

(i.e., to $V_k^{A,F'^M} < V_k^{A,F^M}$). In order to compensate for such a decrease in utility/profits, the insurance provider can respond with a corresponding increase in the private fee (F'^P) as long as the marginal benefit of doing so is greater than its marginal cost. This response was mapped out in the reimbursement-setting process discussed above.

With respect to the second effect, the increase in the Medicare reimbursement increases the minimum private fee required to keep the physician from dropping its current patient k, where $F'^P_{min,j} > F^P_{min,j}$. Thus, it is possible that the current private fee F^P_j is too low to maintain their patient k on the physician's client list. Thus, as long as the insurance provider's utility/profits are greater under a newly optimal private fee than the disagreement utility/profit (which we normalize at zero), then the insurance provider will want to up its private fee offer.